Can AI become more ethical than humans?

Can AI become more ethical than humans?:
A Cross-Paradigmatic Evaluation of the Question

by

Michael Cannon

A thesis submitted for the degree of

Doctor of Philosophy

Spring 2024

Eindhoven University of Technology

Department of Industrial Engineering & Innovation Sciences

Can AI become more ethical than humans?

A Cross-Paradigmatic Evaluation of the Question

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr. S.K. Lenaerts,

voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 25 Juni om 13:30 uur

door

Michael Cannon

geboren te Cambridge, Verenigd Koninkrijk

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	
1e promotor:	Prof. Dr. V. C. Müller
copromotor(en):	Dr. E. O'Neill
leden:	Dr. P. Nickel
	Prof. dr. ir. Jan Broersen (Universiteit Utrecht)
	Prof. Dr. A. Newen (Ruhr-Universität Bochum)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

Abstract

The question of this thesis, whether AI can become more ethical than humans, emerges from literature on artificial superintelligence. This literature is moved by the possibility and plausibility of AI surpassing human capacities. The thesis claim here is that responses to the question of whether AI can become more ethical than humans involve fundamental assumptions about the nature of human and machine cognition, and whether they are the same thing. Recognising that there is not one, but rather at least two paradigms of mind research, this thesis finds that how one responds to the question will depend on which paradigm one adopts and that, so far, existing research and literature on AI in this regard is exclusively of one paradigm.

The paradigms are referred to here as "cognitivism" and "post-cognitivism". Cognitivism is the paradigm of assumptions about mind and cognition upon which existing thinking about AI is based. It is the source of the notion of artificial superintelligence and is identifiable as a paradigm by its operative metaphor that the mind is a particular kind of machine – a computer. In this paradigm, cognition and intelligence are understood as problem-solving capacities, something that emerges from the model of cognition as computation.

By contrast, post-cognitivism conceives of the mind as a living system – an organism – subject to biological and ecological demands. In the same way that the cognitivist model of the mind as a computer leads to computation and problem-solving, post-cognitivism's assumption that the mind is a living system leads to a view in which cognition is the matter of an organism making sense of itself and the world it perceives in order to stay alive.

For the purpose of distinguishing what AI becoming "more ethical than humans" means in each paradigm, this dissertation characterises each paradigm based on their fundamental conception of cognition – in the cognitivist paradigm, cognition is characterised as "problem-solving" because the "problem" itself is taken as "already-given", and in the post-cognitivist paradigm cognition is characterised as "problem-defining" because discerning the problem is included as part of the process of cognition.

The conclusion advanced in the thesis is a conditional: if we think with cognitivism, it is possible that AI can become more ethical than humans – with certain caveats. If we think with post-cognitivism, it is not possible. Beyond this conditional, the broader significance of this conditional conclusion lies in recognising that existing research on AI is based exclusively on cognitivist assumptions about mind and cognition and that, just as it leads to a contingent conclusion about whether AI can be more ethical than humans, it means that AI research more broadly may equally be missing important features when it comes to modelling natural cognition.

1. Introduction

> We can choose who we wish to become when we have decided on an in principle undecidable question.
>
> - Heinz von Foerster

The question of this thesis is "Can AI become more ethical than humans?" The claims of this thesis are twofold. Firstly, an answer depends on the paradigm of mind research with which we think – different existing paradigms express different thinking about the question. Secondly, existing thinking about AI is rooted exclusively in just one of these paradigms and, therefore, may be limited as a model of mind and cognition. Consequently, at the very least, it is unestablished across mind research whether AI can become more ethical than humans, and settling the question will involve discerning how the two paradigms of mind research relate to each other.

This thesis identifies and characterises two existing paradigms of mind research, paradigms which span cognitive science, artificial intelligence, and philosophy of mind. The first paradigm, rooted in Cybernetics, finds deep functional similarities between minds and machines, and is referred to as "cognitivism". This is the paradigm of mind research in which existing thinking about AI takes place. The second paradigm, having developed out of second-order cybernetics and the sciences of living, biological systems, is referred to in this thesis as "post-cognitivism". A chapter is devoted to developing a characterisation of each paradigm. The conclusion of the thesis is that, from the perspective of cognitivism, AI *can* become more ethical than humans, though, it will take more work to determine whether it would necessarily be anything interesting for it may be little more than a moral calculator. From the perspective of post-cognitivism, AI *cannot* become more ethical than humans. Given the opposing conclusions of the respective paradigms, it is significant that existing thinking about AI is an expression of only one of the two paradigms. It means the assumptions of the cognitivist paradigm are consequential as far as ethical AI and AI that is "more ethical than humans" is concerned.

This conditional conclusion emerges from the way I characterise each paradigm. I do so based on distinguishing features of each. I characterise cognitivism by the focus in its model of cognition on "problem-solving" borne of the way in which the model takes the "cognitive situation" as already-given – the agent in question, their problem, and the context in which the process occurs, are variables which are assumed to be well-defined in the cognitivist model. In this model, cognition[1] and intelligence are typically understood in the literature as a process of an agent solving a problem or achieving a goal, notions which will be treated as synonymous here (Legg and Hutter 2007, Russell 2019).

In contrast to cognitivism's "problem-solving", I characterise post-cognitivism by the significance of "problem-defining" in its model of cognition. Rooted in the notion of "sensemaking", the formal conception of cognition in the post-cognitivist literature (Thompson 2007, Froese and Ziemke 2009), I develop an account of "problem-defining" to draw out a contrast between cognitivism and post-cognitivism. As I interpret it, post-cognitivist cognition also involves problem-solving, but is distinguished relative to cognitivism by this "problem-defining" phenomenon.

[1] I use the terms "mind" and "cognition" interchangeably throughout the thesis, acknowledging that they might otherwise be understood differently, with "mind" sometimes being taken to involve phenomenal consciousness. This thesis makes no claims about consciousness and uses the terms synonymously to refer to a cognitive individual, be it organism or artificial computational system.

The distinction between these paradigms and their characterisations is based on the fundamental difference in their basic models of the "cognitive situation". Both models present a situation including an agent, a problem, and an environment to be navigated, but after that, they part ways. The basic but fundamental difference is that the cognitivist model takes the situation for granted, assuming that the situation's variables are already well-defined, and the post-cognitivist model does not. What this means is that the cognitivist model does not account for the genesis and individuation of the very cognitive situation, in effect excluding the dynamics of the cognitive situation's generation from what it models as cognition. The cognitivist model begins with an already-individuated agent, a well-defined problem, and an environment of objectively-defined features to be navigated in order to solve the problem. As Varela et al. put it, in the cognitivist model, the problem in question for an agent is taken as "already-given" (1991: 1). The problem is therefore not in need of discernment or definition.

By contrast, the post-cognitivist model of the cognitive situation includes the dynamics by which problem, problem-solver, and environment come to take on the particular forms and relationships that they do. In this model, the agent endogenously establishes orientation in its situation such that it can endogenously discern the meaning for itself of what it perceives in a way that enables it to keep on living – "problem-defining" is the name I propose for this process. Problem-defining therefore does not mean that the agent is discerning an objective, pre-existing problem external to it. Instead, it simply means that the agent is endogenously discerning for itself, from its perspective, meaning in its situation. This is to say, the meaning of something for an organism, will be the endogenously determined significance it takes on for the organism, from its perspective. I discuss this in several places throughout the thesis noting that it is not intended as a comprehensive account of meaning, but only something which enables me to make the distinctions I want to make. The immediate point here is that, in the post-cognitivist model of cognition in which the genesis of the cognitive situation is accounted for, discerning meaning is an active process. The meaning does not pre-exist the agent. "Problem-Defining" is an attempt to capture this process of meaning-making. Thus, the problem-solving character of cognitivism is fairly simple, but the notion of "problem-defining" requires some development and so I have devoted a section at the beginning of chapter 4 on post-cognitivism to developing the notion.

In the end these respective characterisations are less opposed than their names suggest. The juxtaposition is emphasis for the purposes of the thesis, which is to say, primarily for the sake of distinguishing the paradigms with a clarity that otherwise occludes a nuanced relationship between the two.

These two characteristic conceptions of cognition are, then, the terms in which the question of the thesis is analysed. That is, to the question of whether "AI can become more ethical than humans", the conclusion is that, if becoming more ethical than humans is a matter of AI being better than humans at ethical "problem-solving", then it is possible that AI can become more ethical than humans, because AI are in fact surpassing human performance and capacities at solving well-defined problems. If, however, becoming more ethical than humans involves "problem-defining", determining what the moral problem is, then AI cannot become more ethical than humans because AI is not the kind of entity that is capable of "problem-defining".

The details and path to this conclusion are as follows. The thesis begins with a chapter presenting the point of departure and context of the thesis, namely, the claim that artificial intelligence is an existential risk (Bostrom 2014, Russell 2019). This is the literature in which the thesis question emerged and is the context in which the possibility of artificial intelligence superseding human-level capabilities – intelligence specifically – is developed. The question of whether AI can supersede human-level *morality* is a question raised in Bostrom and Yudkowsky (2014) when they write: "This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes

more ethical than you?" (Bostrom and Yudkowsky 2014: 16). As a field of research exploring the question of how to think about machines as moral agents, this is not in fact a question to which Machine Ethics (Anderson and Anderson 2007, 2011) currently devotes much attention, with the notable exception of (Petersen 2017)[2].

The crucial turn from here in the thesis is the idea that the very concepts involved in the claim that superintelligence is an existential risk are distinctly features of the cognitivist paradigm and its fundamental assumptions. Post-cognitivism has different fundamental assumptions, and from these "initial conditions" is thus led to different ideas about the potential of AI. The third, fourth, and fifth chapters of the thesis present each paradigm in turn to bring this out.

The third chapter introduces cognitivism, bringing out its "problem-solving" character, borne of taking the cognitive situation as already-given. The chapter begins by noting how cognitivism is defined in the literature and then charts a path through three philosophical notions fundamental to it: computation, information, and functionalism. Taken together, these concepts are sufficient to show how deeply the "problem-solving" conception of cognition runs in cognitivism. Ordinarily a notion of "representation" would be considered a necessary feature of cognitivism, but it is excluded from this thesis on the basis of being unnecessary to understand the "problem-solving" character of cognitivism. This does then mean that the description of cognitivism in this thesis is not exhaustive of the paradigm, and yet, the aspiration of the thesis is to highlight something nonetheless distinguishing and fundamental about it.

Within the cognitivist paradigm, there are multiple theories of cognition, ranging from "classical", "connectionist", and even "embodied" (Varela et al. 1991, Thompson 2007). In order to define the three cognitivist concepts of focus – computation, information, functionalism – in a way that is general enough to include this diversity of theories, and detailed enough to avoid triviality, the chapter on cognitivism returns to the fundamental ideas in which the concepts are rooted. For computation, the chapter chronicles Turing's work on computable numbers (Turing 1937) and Turing machines (ibid 1950) before looking to contemporary discussions and definitions of computation in philosophy of mind. For information, most of the space is devoted to talking about the development of information theory (Shannon 1948). Meanwhile, the discussion of functionalism begins with a discussion of the work of Hillary Putnam (1960, 1967, 1975, 1988), the philosopher who first explicitly introduced the idea in philosophy of mind, and then turns to the way functionalism was employed in cognitive science in the work of David Marr's famous Tri-Level Hypothesis for analysing information-processing systems (Marr 2010). The aim of the chapter is, firstly, to show how these fundamental ideas jointly produce a conception of cognition which views it as a matter of an agent solving a problem.

With that established, in the fourth chapter, I put this material to work to show how existing thinking on AI, including superintelligence, is rooted in this particular conception of cognition. From there I argue that the AI existential risk is a set of ideas based distinctly on these cognitivist notions and their assumptions. This is to say, cognitivism is effectively the way we are already thinking about AI.

In chapter five I introduce post-cognitivism and focus on a particular theory of cognition taken to be representative of the paradigm, the Enactive Theory of mind (Thompson 2007, Froese and Ziemke 2009). I begin the chapter with a section defining and developing what I mean by "problem-defining", identifying it as the "endogenous orientation and discernment of meaning" that *precedes* problem-solving. With the notion of problem-defining in place, I then describe the enactivist theory of cognition to show how it has this problem-defining character. I finish the chapter with a discussion of "4E"

[2] In contrast to Bostrom (2012, 2014), Petersen argues that an artificial superintelligence will reason about what its final goals are, and that "this is where ethics can get a foothold" (Petersen 2017: 2).

cognition, the broader theory in which Enactivism is embedded. 4E cognition is a model of cognition which holds that the mind is embodied, embedded, extended, and enactive (Aizawa 2018, Newen et al. 2018). Following the work of Evan Thompson in *Mind in Life* (2007), the four concepts fundamental to the enactive theory are the life-mind continuity thesis (Kirchoff and Froese 2017), autopoiesis (Maturana 1970, Maturana and Varela 2012), sensemaking, and the dynamic co-emergence of self and world (Thompson 2007). Together these concepts produce a picture of cognition quite distinct from the cognitivist's computational picture.

The enactive theory of mind identifies individuals as autopoietic, which means "self-producing". This is the starting condition from which the conception of enactive cognition develops. In the course of autopoiesis, an agent perceives and reacts in a way that is adaptive, and this condition of existence defines the basic cognitive problem and situation: the autopoietic individual must make sense of perturbations to themselves in a way that keeps them alive – in this way they must establish and continuously re-establish their orientation and basis for discernment of meaning, that is, problem-defining[3].

In contrast to this endogenous, autopoietic individuation, agents in the cognitivist paradigm are usually, and usually implicitly, defined by an external observer, engineer or AI researchers and so on. The paradigm emerged out of the development of computational machines (Gleick 2012, Dupuy 2000), both abstract, like Turing machines, and concrete, like the telegrams, telephones, anti-aircraft weapons systems and other electrical engineering systems on which people like Claude Shannon (1948) and Norbert Wiener (1948) worked. These machines were designed and arranged by engineers for specific problems, problems belonging to the humans who built them. The machines were built to solve those problems. In the course of a history of science and culture well documented by James Gleick (2012) and Jean-Pierre Dupuy (2000), such machines become powerful models for the workings of the human minds that made them. The key is, the model does not include how the cognitive agent is generated in the first place, but, it should be said, this stands to reason – we have specific uses for telephones and, when we model that use and process, we do not need to include the genesis of the telephone. By contrast, when it comes to modelling the minds of humans – humans who happen to be trying to model their own minds – we *are* dealing with something that is trying to model its own process and therefore the genesis is a relevant part to be included in the model, should we wish it to be considered complete.[4]

With these distinctions in place, the last chapter presents the conclusion of the thesis in the form of two conditional statements. On the one hand, if we think with cognitivism about whether AI can become more ethical than humans, then, as long as the problem of ethics is already defined, AI might be able to become more ethical than humans in the sense that it can better solve moral problems. What the moral status of such a system would be is unclear, and, further, it is not clear to me that this would be an interesting sense of "more ethical" because such a system might be little more than a moral calculator, with as much significance in ethics as calculators have in mathematics. On the other hand, if we think

[3] An important caveat that will be discussed is that autopoiesis does not unanimously entail cognition for theorists in the post-cognitivist paradigm (Thompson 2007: 122-127).

[4] This tracks the distinction between 1st and 2nd Order Cybernetics (Dupuy 2000, von Foerster 2005). 1st Order Cybernetics, or just 'cybernetics', is concerned with systems which exhibit goal-oriented, "purposive" behaviour (see (Rosenblueth et al. 1943) for a seminal paper). With a feedback loop to regulate its behaviour, a telephone counts as a "non-teleological" system, whilst mechanical systems with feedback loops are "teleological". Where cybernetics is concerned with goal-oriented systems, 2nd Order Cybernetics is concerned with humans building cybernetic systems – the "cybernetics of cybernetics" (von Foerster 2005). A fuller account of cognitivism and post-cognitivism than is possible in this thesis would have included their respective heritage in cybernetics and 2nd order cybernetics.

with post-cognitivism, AI cannot become more ethical than humans because this would require defining what the "problem" is, and that is not something AI can do yet.

In the final chapter, I summarise the arguments of the thesis in the form of a series of premises and conclusions. The main conclusion of the thesis is represented by C1.

> **C1: The two paradigms of mind research, Cognitivism and Post-Cognitivism, lead to different conclusions about whether AI can become more ethical than humans.**

This is supported by two supporting conclusions concerning the respective positions of the cognitivist and post-cognitivist paradigms on the question of whether AI can become more ethical than humans.

> **C2: If we take a cognitivist approach to cognition, we can be led to the conclusion that AI *can* become more ethical than humans.**
>
> **C3: If we take a post-cognitivist approach to cognition, we will conclude that AI *cannot* become more ethical than humans.**

The case for C2 is made in chapter 3, and the case for C3 is made in chapter 4. I explicitly summarise the premises leading to these conclusions in chapter 5. Again, the significance of these divergent and conditional conclusions lies in the fact that existing thinking about AI makes exclusively cognitivist assumptions about cognition. Because these assumptions are not universally shared across mind research, this prompts questions in at least two regards. Firstly, it means we need to ask how seriously we should take speculations about the scaling of intelligence to artificial general intelligence and artificial superintelligence, which depend on particular assumptions about cognition and intelligence. More broadly, if those assumptions are appropriate for accounting for the operation of machine systems, it brings into question whether this necessarily means such machine systems are legitimate models of human and natural cognition, or whether they are missing something by making the assumptions that they do.

The existence of different paradigms does not in itself prove anything either way, but existing thinking about AI does not yet pay attention to post-cognitivist work. In this thesis, I aim to offer a first step in this direction.

2. Context: Existential Risk, Intelligence and Artificial Superintelligence, the Orthogonality Thesis

The title question of this thesis, "can AI become more ethical than humans?", is directly taken from the paper "The Ethics of Artificial Intelligence", written by Nick Bostrom and Eliezer Yudkowsky (2014). The paper is written in the context of the discussion about the existential risk of artificial superintelligence (ASI), for which Bostrom (2014) is taken as the seminal text in philosophy. In this context, the possibility of ethical superintelligence is discussed as a strategy for mitigating the risk of ASI to humanity. The idea is that, if a superintelligence is ethical, it is less likely to be a danger to us. Of course, this is no guarantee. There may be good moral reasons to get rid of humans. In any case, some also believe superintelligence could be a good thing, enabling and empowering humanity toward flourishing futures. The question of ethical artificial intelligence is not just a matter of risk-mitigation to avoid bad outcomes, but also concerns actively enabling positive outcomes.

With regard to the question of whether AI can become more ethical than humans, this literature on the existential risk of superintelligence represents the point of departure for this thesis. There are already many discussions of the ways in which, and the extent to which, artificial systems might be ethical, For example, there exist discussions of "artificial moral advisors" (Savulescu and Maslen 2015; Giubilini and Savulescu 2018), "robot rights" (Gunkel 2012, 2018), "robot ethics" (Nyholm 2020, Coeckelbergh 2022), "Robophilosophy" (Seibt et al. 2014, Hakli and Seibt 2017, Seibt 2020) and "machine ethics" (Floridi and Sanders 2004, Moor 2006, Anderson and Anderson 2007, 2011, Brundage 2014, Wallach and Allen 2008, Wallach and Asaro 2017). Each of these discussions obviously have much to offer to a discussion of AI becoming more ethical than humans, however, it is important to begin with the discussion about ASI for at least two reasons.

The first is simply that the literature on superintelligence is the most developed discussion on the possibility of AI comprehensively *superseding* humans in some capacity, and so is particularly relevant to the question of AI becoming more ethical than humans. Supersession is not a central concern of the other discussions in philosophy of AI. The literature on superintelligence even speaks relatively directly to the question AI becoming more ethical than humans. The "Orthogonality Thesis" (Bostrom 2012: 74, 2014: 107) in particular has major implications for whether artificial superintelligence would also be super-ethical, or ethical at all. It makes for a fruitful starting point on this alone.

The second reason is more fundamental. The claim that artificial intelligence may become superintelligent and thereby an existential risk, is a claim rooted in a particular conception of mind, referred to in this thesis as "cognitivism". Recognising this is key to appreciating the significance of the conclusions of this thesis. In the second and third chapter I explore the way in which the existential risk of AI is fruit of the cognitivist ecosystem of thought in particular. The motivation for this discussion is that the existential risk claim is quite a serious claim, and some of the ideas necessary to support it, like the Orthogonality Thesis, are very counterintuitive. In such a situation, one path is to accept that things are a bit weird, look for explanations which assuage our disgruntled intuitions and carry on. Another path is step back and reconsider the foundations by which we came to be in the situation.

One challenge in the context of ethical AI is that such foundations are rarely discussed, let alone explicitly presented as assumptions or antecedents. This may be to do with the way in which ideas and research flow between philosophy, the cognitive sciences, and AI research. In a paper discussing "new developments in philosophy of AI" (Müller 2016) makes the analogy that cognitive science and AI research have long been married, a union facilitated by philosophy as the "best man" (the shared

philosophical assumptions in cognitive science and AI), but that the two spouses are becoming more independent, or not "talking" as much. To the extent that conversation and commerce between the domains is less than it could be, it is conceivable that the assumptions are not subject to the same kind of exposure and analysis as they might have once been.

Thankfully, though the foundations are not often discussed, they are not hidden either. The antecedent assumptions are particularly prominent in discussions of superintelligence, where they are taken to their logical extremes. In the third chapter, I will explore the way in which superintelligence itself can be understood as the logical conclusion of a set of "cognitivist" foundational assumptions about "mind". More broadly than superintelligence and existential risk, the cognitivist conception of cognition I outline in the chapter is the paradigm of assumptions in which historical and existing thinking about AI is exclusively based.

Whilst these foundations are well supported both theoretically and empirically in cognitive science and technical artificial intelligence research, they are not universally shared, which means other sources of response to the question of this thesis are possible, not to mention other ways of thinking about AI. In the fourth chapter of this thesis, on the Enactive and "4E's" account of mind and cognition, I present a characterisation of the "post-cognitivist" paradigm of mind, a paradigm developed in major sources like (Varela et al. 1991, Thompson 2007, Newen et al. 2018). This paradigm of mind has quite different fundamental assumptions. Existential risk is not a concern of this paradigm.

In short, discussion of superintelligence affords a certain exposure to our existing biases in thinking about AI. Given a measure of such exposure, it becomes clear that there are alternative possible sources of answers to the question of this thesis and thinking about AI generally – because what one thinks about AI is necessarily informed by what one thinks about minds. A sensibility for the various ways we can think about the question in turn enables a more comprehensive response to the question. A more self-aware and less arbitrary philosophy of AI and mind becomes possible when we can choose between paradigms.

There is one further, "addendum" reason for which the literature on superintelligence makes for a good point of departure. For those that take the proposition sincerely, there is something real-world at stake in understanding the "minds" of artificial intelligence. For those for whom there is something at stake, the goal of the field is not just explanatory, as it mostly is with the mind sciences more broadly, but nothing less than to protect and preserve civilisation. This stake has a generative way of clarifying specific questions relevant to research and is most present in the literature on superintelligence.

Here is the passage from Bostrom and Yudkowsky which inspired the thesis:

> *"This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes more ethical than you?* This is not like asking our own philosophers to produce superethics, any more than Deep Blue was constructed by getting the best human chess players to program in good moves. *But we have to be able to effectively describe the question, if not the answer*—rolling dice won't generate good chess moves, or good ethics either. (Bostrom and Yudkowsky 2014: 16-17) [emphasis mine]

There are at least two things that are striking about this passage. The first is simply the provocation of the proposition of AI becoming more ethical than you. The second is the way it directs attention to the fact that such a proposition requires understanding the nature of the question before any answer. I want

to go a little further - in a domain that is still largely uncharted, it seems important that we take adequate time to discern what the questions *might be* and are open to as many possible understandings as we can be. If nothing else, it seems at least important to hold space for changing what we take the question to be. The risk is that we settle on a path of inquiry and are led down a path of conceptual affordances which trap us into certain ways of thinking about AI which "becomes more ethical than you". Foundational work often involves holding in question just where these paths of inquiry are or should be. In the language of this metaphor, this thesis is comparing two such paths and in-principle remaining ambivalent about which we should take. The aim of this thesis is ultimately an attempt to understand what it might mean for AI to be "more ethical than humans" and to define it in a way that opens up new territories of inquiry. Answers can be thought-terminating, the way programs halt upon successful execution of a task, so the aim in this thesis is to be thought-generating in this regard.

Having said that, whilst this thesis is in-principle ambivalent, one of the main points of this is that, in-fact, existing thinking about AI is exclusively of just one of the paradigms and therefore relies on a particular set of assumptions that are not universally shared across mind research. This means that those assumptions should not be taken for granted but should be open to investigation. Whilst post-cognitivist theories often establish themselves by first emphasising their different foundational assumptions in relation to cognitivist theories, cognitivist work in AI, particularly on the matter of superintelligence, does not yet seriously engage its own assumptions about cognition and intelligence. My work in this thesis is therefore to take a step in this direction and show that these assumptions do make a difference when it comes to thinking about AI and whether it can be more ethical than humans.

The literature on superintelligence therefore stands as a worthy starting point and context for an investigation into the question of this thesis. I will now turn to present the current conception of superintelligence in the context of existential risk of artificial intelligence in order to highlight what this literature has to say about whether AI could become more ethical than humans. The paradoxes in the literature make clear the extent to which it seems to be the product of "cognitivism".

2.1. Superintelligence, XRisk, and Orthogonality

There are three concepts from the literature on ASI which are sufficient to understand the picture. Perhaps the most important initial point is that the proposition of AI becoming more ethical than humans has emerged from discussion of the proposition of AI becoming more *intelligent* than humans. The question of *artificial superintelligence* (superintelligence), the first concept, came prior to any question of "artificial superethics". Discussion of superintelligence mostly occurs in the context of discussions about the *existential risk* (xrisk) it poses. *Xrisk* is the second concept. The concern is that superintelligence could have different goals to humans and that, were humans to come into a rivalrous relationship with such a superintelligence, given the AI is by definition far more intelligent than humans, this is a rivalry humans would probably lose. And it may not be out of malevolence that a superintelligence ended humanity. In pursuit of its idiosyncratic goals, it may be for instrumental reasons and instrumental goals that a superintelligence ended humans with indifference. The belief that superintelligence could have very different goals to humans emerges from a paradoxical, but disarmingly simple and intuitive thesis known as the *Orthogonality Thesis*, the third and final concept.

These three concepts – xrisk, superintelligence, and orthogonality – form the conceptual backbone of the literature on superintelligence, of which the question of "superethical" AI is a branch. I will therefore now define and present the concepts as they are discussed in the literature and then turn toward how we

might "be able to effectively describe the question, if not the answer" (Bostrom and Yudkowsky: 16-17) of whether AI can become more ethical than humans. The Orthogonality Thesis is where most of the action really happens, so more time is devoted to it.

2.1.1. Existential Risk

Existential risk is the simplest of the three concepts here and doesn't require much explanation for present purposes. Bostrom defines "xrisks" as those risks which "threaten the entire future of humanity" (Bostrom 2013):

> "An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development." (Bostrom 2013: 1; see also 2002)

Bostrom has been writing about large-scale risks for a while, including a co-edited book *Global Catastrophic Risk* (Bostrom and Cirkovic 2011). Global catastrophic risks and xrisks are the same kind of thing, differentiated by degree, with xrisk being the most extreme. The categorization is understood as the product of at least two variables – scope and severity – as visually presented on the table below (Fig. 1) from (Bostrom 2013: 17)

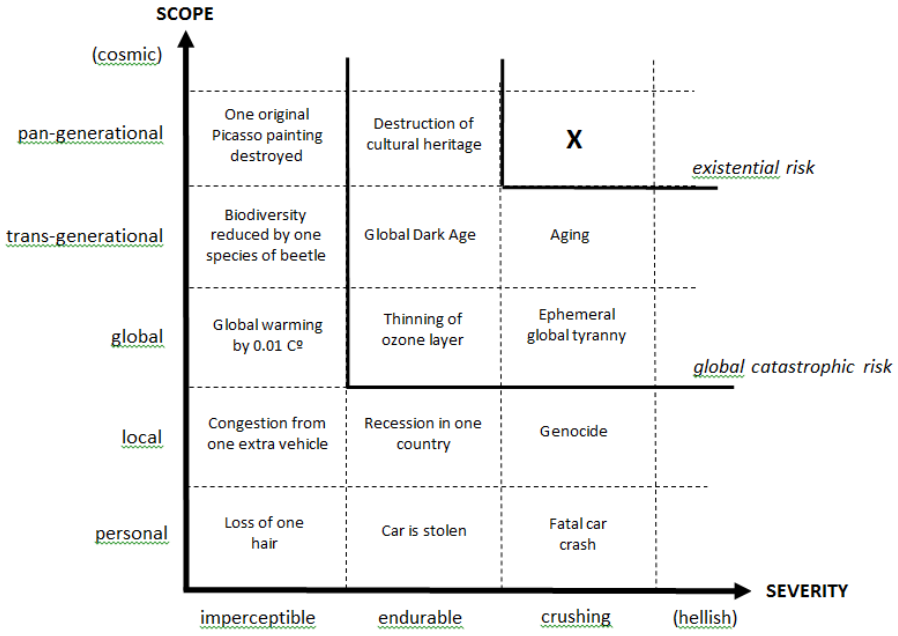


Figure 1. Qualitative Risk Categories from (Bostrom 2013)

It is in this existential risk category, at the intersection of "hellish" and "pan-generational" values on the axes, that artificial superintelligence is put. We can now explore this notion of superintelligence.

2.1.2. Intelligence and Artificial Superintelligence

The now famous first articulation of what superintelligence is comes from Irving Good's "Speculations concerning the first ultraintelligent machine", in a section titled "Ultraintelligent Machines and their Value":

> "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever." (Good 1966: 33),

The most (in)famous contemporary work on superintelligence is Nick Bostrom's (2016) *Superintelligence: Paths, Dangers, Strategies*. (**In**)famous because there are some academics for whom, at least anecdotally, the notion of Superintelligence is nonsense, e.g. (Mitchell 2019). I have included the prefix therefore to acknowledge the relative controversy. Part of the challenge in discussing views on superintelligence is that often authors who do not agree with the idea do not spend time critiquing it and just get on with their own work. In any case, Bostrom defines superintelligence very similarly to Irving Good, even citing in a footnote to this quote the heritage and similarity with Good's definition:

> "We can tentatively define a superintelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.*" (Bostrom 2014: 22)

It is a tentative definition insofar as it says little more than "that thing humans have, it's like that, but more of it, much more of it." It does not suffer from being unintuitive at least, but it does leave open the question of whether it matters what is meant by "intellect" or "intelligence". The elephant in the room in Bostrom's *Superintelligence*, and associated discussions (in particular Omohundro's work 2007, 2008, 2012), and also more recent work like Russell (2019), is the paucity of space and attention given to discussing just what "intelligence" is. Naturally, it is just plain difficult to do justice to a phenomenon as enigmatic as "intelligence", let alone have any space for talking about anything else after having attempted to do so, so this particular point of critique is there simply to highlight that the conclusions which Bostrom and others reach are products of where they have chosen to draw the line with respect to defining intelligence. Different conclusions might be reached with different antecedent ideas about "intelligence". I will discuss this further later on, but it is important to highlight it now so that the logical trajectory of ideas is retained as I present them.

2.1.2.1. Intelligence as Instrumental Rationality

Where intelligence is discussed, the summary conception of it is as "instrumental rationality". In one of the few places where Bostrom (2014) explicitly addresses the question of "intelligence", he characterises it casually, but intuitively:

> "By "intelligence", we here mean something like skill at prediction, planning, and *means-end reasoning* in general. This sense of *instrumental cognitive efficaciousness* is most relevant when we are seeking to understand *what the causal impact of a machine superintelligence might be*." (Bostrom 2014: 107) (emphasis mine)

Bostrom's concerns here are strategic. He is not primarily concerned with a conceptual analysis of intelligence, he is concerned with machines developing a certain level of capacity, a capacity that can be sufficiently described by a notion of intelligence as a means-end capacity. His notion of intelligence therefore stands as both conceptual analysis and practical work. In his earlier work, (2012) Bostrom defines intelligence similarly to above:

> For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning … Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. (Bostrom 2012: 73)

Both definitions draw attention to the *instrumental* service of the reasoning. This point is particularly important in the context of the Orthogonality Thesis, which I discuss next, for the Orthogonality Thesis follows from such a conception of intelligence.

Bostrom's account here is consistent with, and informed by, several papers by Stephen Omohundro (2007, 2008, 2012). These papers speculate about the nature of self-improving AI and the way in which it could become an xrisk. In an endnote, Bostrom describes these as "pioneering papers on this topic" (Bostrom 2014: 280, endnote 8). Bostrom's and Omohundro's account converge in many respects, and in particular with respect to the instrumental, means-end nature of the intellectual capacities of AI:

> To say that a system of any design is an "artificial intelligence", we mean that it has goals which it tries to accomplish by acting in the world. If an AI is at all sophisticated, it will have at least some ability to look ahead and envision the consequences of its actions. (Omohundro 2008: 1-2)

One of the particularities of Omohundro's account, consistent throughout his (2007, 2008, 2012) papers, is the claim that AI will "converge on rational economic behaviour", behaviour formally defined by the norms of expected utility theory. Omohundro suggests that in order to stay alive and "self-improve" (2007) entities will behave in predictable ways given their goals, and that these patterns of behaviour are describable in terms of the concepts of rational economic behaviour, concepts like "goal" and "utility" with respect to that goal.

Another important author and paper in this space is Stuart Armstrong's (2013) paper. He expresses a similar, and similarly pragmatic, account to both Omohundro and Bostrom.

> "But even 'intelligence', as generally used, has too many connotations. A better term would be efficiency, or instrumental rationality, or the ability to effectively solve problems given limited knowledge and resources... Nevertheless, we will be sticking with terminology such as 'intelligent agent', 'artificial intelligence' or 'superintelligence', as they are well

> established, but using them synonymously with 'efficient agent', artificial efficiency' and 'superefficient algorithm'. The *relevant criteria* is whether the agent can effectively achieve its goals in general situations, not whether its inner process matches up with a particular definition of what intelligence is." (Armstrong 2013: 69) (emphasis mine)

The pragmatic orientation of these authors with respect to the conception of intelligence is perhaps most explicit in (Yudkowsky 2013):

> "… we may be better off with "Intelligence is that sort of smartish stuff coming out of brains, which can play chess, and price bonds, and persuade people to buy bonds, and invent guns, and figure out gravity by looking at wandering lights in the sky; and which, if a machine intelligence had it in large quantities, might let it invent molecular nanotechnology; and so on." To frame it another way, if something is powerful enough to build a Dyson Sphere, it doesn't really matter very much whether we call it "intelligent" or not. And this is just the sort of "intelligence" we're interested in – something powerful enough that whether or not we define it as "intelligent" is moot…Choice of definitions has no power to effect physical reality." (Yudkowsky 2013: 9)

It helps to understand that these authors are less interested in a complete conceptual analysis of the concepts they employ, and more in a figuring out how to solve a problem. It makes sense of this kind of thinking. That said, their conception of intelligence is robust and intuitive. They are each at pains to avoid anthropomorphising intelligence by insisting that certain features which we, being humans, might presume to be fundamental and integral to intelligence, need not in fact be, and that intelligence can be quite unlike our own. There is much more to say about this, and it will be discussed in further detail in the next section on the Orthogonality Thesis. I raise it here to allay concerns that the pragmatism may be a bit overdone or even an abdication of deeper inquiry.

The most contemporary discussions in this context are Stuart Russell's (2019) book *Human Compatible: AI and the Problem of Control* and Brian Christian's (2020) *The Alignment Problem: How can machines learn human values?* Of these two, Christian does not explicitly devote any discussion to the notion of intelligence, citing Legg and Hutter (2007a, 2007b) and Legg and Veness (2011) "for more on the idea of a universal definition of intelligence" (Christian 2020: 150, footnote 87). Legg and Hutter are discussed below. They are the go-to source on intelligence for many authors in this space. Beyond this reference, and speaking of the role of reinforcement learning in the development of AI, Christian does note the limits of a notion of merely instrumental intelligence:

> "Reinforcement learning also offers us a powerful, and perhaps universal definition of what intelligence *is.* If intelligence is, as computer scientist John McCarthy famous said, "the computational part of the ability to achieve goals in the world," then reinforcement learning offers a strikingly general toolbox for doing so. Indeed, it is likely that its core principles were stumbled onto by evolution time and again – and it is likely that they will form the bedrock of whatever artificial intelligences the twenty-first century has in story.

> In some ways, though, a deeper understanding of the *ability* of animals and machines to achieve goals in the world has kicked the more profound philosophical can down the road. This theory, pointedly, does not tell us *what* we value, or what we *ought* to value." (Christian 2020: 150-151) (emphasis in original)

Russell gives the notion of intelligence comparatively more space, devoting the whole of the second chapter "Intelligence in Humans and Machines" to establishing an informed and robust account of intelligence which can serve as foundation for the rest of the book. That said, he nonetheless articulates effectively the same account, shifting in terms from instrumental, means-end reasoning to *rationality* more generally. The difference between means-end reasoning and rationality in his sense is not significant. The other authors mentioned above seem to intend much the same thing which Russell describes.

Russell begins similarly to the others: "From the earliest beginnings of ancient Greek philosophy, the concept of intelligence has been tied to the ability to perceive, to reason, and to act *successfully*." (Russell 2019: 20) Echoing Bostrom, Russell goes on to cite Aristotle (Nicomachean Ethics, Book III, 3 1112b) to support the idea that intelligence is oriented with respect to means-end, instrumental reasoning (ibid). Russell then supplements his quoted passage from Aristotle with a modern update to rationality in which means-end reasoning includes reasoning about uncertainty, the norms for which are described by expected utility theory (Russell 2019: 20-27). "Rationality" subtly becomes a heuristic for "intelligence", which is understandable at least though, insofar as rationality is far more "well-defined" phenomenon than intelligence:

> "In short, *a rational agent acts so as to maximize expected utility.* It's hard to overstate the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines." (Russell 2019: 23) (emphasis in original)

Following this description of "rationality for one", Russell goes on to describe "rationality for two" in which the norms of means-end reasoning are adapted to account for other agents, as per game theory (Russell 2019: 27-32). What remains is that the norms of reasoning still assume the nature of the problem, including the relevant ends or goals, and amount to a deliberation about means to reach those ends. This is to say, intelligence can still be understood as *instrumental rationality.*

So far, in this conception of intelligence as something like the rational pursuit of ends, intelligence can be understood as a navigation problem of sorts, the problem of navigating a possibility space in order to arrive at the desired end. Taking this a bit further, some authors have suggested understanding this instrumental rationality as an optimisation process, a matter of finding the optimal path through a space of possible paths to a goal:

> "The notion of an "optimization process" is predictively useful because it can be easier to understand the target of an optimization process than to understand its step-by-step dynamics." (Yudkowsky 2008: 10)

This is an "effective" way of conceiving of the "intelligence" of those AI systems whose "step-by-step dynamics" are in some way impenetrable to human, whether for "black box" reasons, or for human-limits reasons. We nonetheless still understand that the system is optimising for a particular goal, and that each of its "steps" are instrumentally rational in this regard.

Roman Yampolskiy has thought and written extensively about superintelligence, most notably in his (2016) book *Artificial Superintelligence: A Futuristic Approach*. It will turn up again in the next section on Orthogonality. In a separate paper, preceding his book, he and his co-author also employ the concept of optimisation to articulate their conception of "mind", the ontological unit, so to speak, usually understood in terms of being the bearer of the properties like intelligence:

> "We use the term "mind" here simply as a synonym for an optimizing agent. Although the concept "mind" has no commonly-accepted definition beyond the human example, in the common intuition, humans and perhaps some other higher-order animals have a mind. In some usages of the term, introspective capacity, a localized implementation, or embodiment may be required. In our understanding, any optimization process, including a hypothetical artificially intelligent agent above a certain threshold, would constitute a mind. Nonetheless, the intuitions for the concepts of "mind" and "intelligence" are bound up with many human properties, while our focus is simply on agents that can impact our human future. For our purposes, then, the terms "mind" and "intelligence" may simply be read "optimizing agent" and "optimization power." (Yampolskiy and Fox 2012: 3)

The shift from instrumental reasoning to instrumental rationality to optimisation process makes for a shift toward increasingly austere, functionalist conceptions of intelligence which increasingly remove any "mentalistic" or psychological concepts from their description in favour of mechanistic and behaviouristic descriptions. This trend is indicative of the cognitivist heritage in behaviouristic psychology (Dupuy 2000). However intuitive we find any of them, and whether we agree with them or not, it is important to note that this picture is very much at feature of the "modern" paradigm for building artificial intelligence, the paradigm of "intelligent agents", which is to say, "rational agents". (Russell and Norvig 2010, Russell 2019: 42).

Further, this conception of intelligence is not isolated to work on superintelligence or even AI but is in fact consistent with definitions of intelligence across domains of research, as per Legg and Hutter's (2007) paper. Their paper synthesised over 70 scholarly definitions of "intelligence" from across several domains, producing this result:

> "Intelligence measures an agent's ability to achieve goals in a wide range of environments." (Legg and Hutter 2007: 9)

Read in the terms of the conception of intelligence as instrumental rationality discussed above, Legg and Hutter's synthesis can be reinterpreted along the lines of "intelligence measures an agent's instrumental rationality or optimisation ability to achieve goals in a wide range of environments".

In any case, there is in the literature an awareness of the limits, dangers even, of an unthinking application of this account of intelligence. Russell notes the limits of this "standard model" for the way

in which it is effective for designing machines which can be expected to pursue their goals, but not effective for designing machines which can be expected to pursue our goals. This foreshadows the discussion of the Orthogonality Thesis. Russell writes:

> "So, how has the field of AI approached the "design machines with a high degree of intelligence" part of the task in the past? Like many other fields, AI has adopted the standard model: we build optimizing machines, we feed objectives into them, and off they go. That worked when the machines were stupid and had a limited scope of action; if you put in the wrong objective, you had a good chance of being able to switch off the machine, fix the problem, and try again.
>
> As machines designed according to the standard model become more intelligent, however, and as their scope of action becomes more global, the approach becomes untenable. Such machines will pursue their objective, no matter how wrong it is; they will resist attempts to switch them off; and they will acquire any and all resources that contribute to achieving the objective." (Russell 2019: 172)

Having pointed out the way in which intelligence is generally understood in terms of instrumental rationality, I will now return to add a few more details to the picture of superintelligence itself, as per Bostrom (2014). In his third chapter he considers three different forms of superintelligence:

> "As previously indicated, we use the term "superintelligence" to refer to intellects that greatly outperform the best current human minds across many very general cognitive domains. This is still quite vague. Different kinds of systems with rather disparate performance attributes could qualify as superintelligences under this definition. To advance the analysis, it is helpful to disaggregate this simple notion of superintelligence by distinguishing different bundles of intellectual super-capabilities. There are many ways in which such decomposition could be done. Here we will *differentiate between three forms: speed superintelligence, collective superintelligence, and quality superintelligence.* (Bostrom 2014: 52) (emphasis mine).

I want to draw attention to the words "disaggregate", "bundle", and "decomposition" for the way in which they speak of an aggregation of modular cognitive functions. This kind of thinking will turn out to be central in the discussion on Orthogonality. Addressing each form of superintelligence in turn, Bostrom says the following:

> "A speed superintelligence is an intellect that is just like a human mind but faster. This is conceptually the easiest form of superintelligence to analyse. We can define speed superintelligence as follows:
>
> **Speed superintelligence:** *A system that can do all that a human intellect can do, but much faster.*
>
> By "much" we here mean something like "multiple orders of magnitude." But rather than try to expunge every remnant of vagueness from the definition, we will entrust the reader with interpreting it sensibly.

> The simplest example of speed superintelligence would be a whole brain emulation running on fast hardware. An emulation operating at a speed of ten thousand times that of a biological brain would be able to read a book in a few seconds and write a PhD thesis in an afternoon. With a speedup factor of a million, an emulation could accomplish an entire millennium of intellectual work in one working day." (Bostrom 2014: 53) (emphasis in original)

Bostrom footnotes some caveats and details which are worth considering, but unnecessary for my purposes here. The next form of superintelligence he discusses is "collective superintelligence":

> "**Collective Superintelligence:** A system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system.
>
> Collective superintelligence is less conceptually clear-cut than speed superintelligence. However, it is more familiar empirically. While we have no experience with human-level minds that differ significantly in clock speed, we *do* have ample experience with collective intelligence, systems composed of various numbers of human-level components working together with various degrees of efficiency." (Bostrom 2014: 54) (emphasis is in original)

Teams, organisations, companies, societies and civilisation are among the examples of collective intelligence which Bostrom presents for our intuitions. He is quiet about whether there is any significant distinction between something like an *artificial* collective superintelligence, which presumably would be some systemic integration of many AI or superintelligent systems, and those cited examples of aggregates of humans.

The third form of superintelligence that Bostrom presents is "quality superintelligence":

> "**Quality Superintelligence:** A system that is at least as fast as a human mind and vastly qualitatively smarter." (Bostrom 2014: 56) (emphasis in original)

Bostrom admits that this is the most ambiguous of the forms which he presents, and points to the difference between human and nonhuman/animal minds for a reference and sense for the measure to which he is trying to point.

> "A zebrafish has a quality of intelligence that is excellently adapted to its ecological needs; but the relevant perspective here is a more anthropocentric one; our concern is with performance on *humanly* relevant complex cognitive tasks…
>
> …the concept of quality superintelligence: it is the intelligence of quality at least as superior to that of human intelligence as the quality of human intelligence is superior to that of elephants', dolphins', or chimpanzees'." (Bostrom 2014: 56-57) (emphasis in original)

This is a hard one to précis. It deserves a little more detailed discussions than the others.

Perhaps the simplest way of putting it is to do with the way in which "intelligence" operates as a loose ordering principle for basic differences in capacities in the animal kingdom, differences which we tend to render and order in a hierarchy of organisation. Colloquially we discern a difference between single-celled organisms and bugs, and between bugs and rodents, and rodents and chimpanzees, and chimpanzees and humans, and so on, and this particular kind of difference appears to be on the same qualitative and vertical scale, which is to say, it seems to be a measure of the same thing. Despite the consistent failures of our collective efforts to discern and define that quality in a non-arbitrary way, it remains intuitive to name it "intelligence", though some do point to the ways in which it is more nuanced than this (Bhatnagar et al. 2018). Halina (2015) also offers a dissenting opinion on this "Standard Model" of intelligence[5]:

> "Rather than attempt to categorize the natural world into organisms that are intelligent on the one hand and rely on rigid, mechanical processes on the other, we should focus on specific mechanisms and behaviours that take an organism's evolutionary, developmental, and ecological context into account. The term "intelligence" should be recognized as an umbrella term that orients researchers towards a broad class of phenomena but is too general for detailed scientific work." (Halina 2015: 2)

To the extent that the "standard model" is the picture Bostrom has in mind, the concept of "quality superintelligence" can be imagined as on this same scale, on which cells, bugs, rodents, apes are ordered, with the superintelligence being to humans as humans are to, say, "elephants, dolphins, or chimpanzees".

This is an intuitive way to conceive of intelligence. Whilst Bostrom does seem to mean "quality superintelligence" in this way, he is also a bit more specific too, detailing an interpretation of quality superintelligence which is consistent with the above account of intelligence as "instrumental rationality". Where he mentions above the ecologically well-adapted capacities of the zebrafish, intelligence is still conceived in terms of capacities to perform certain tasks. Intelligence is still conceived of in terms of an instrumental capacity to achieve ends, however species-relative and domain-specific those ends and their requisite capacities may be.

It would be easy and intuitive to suppose of something like "quality superintelligence" that it meant an intelligence that was in service to better quality *goals*, more significant goals perhaps, or at the least in service to the discernment of better-quality goals, but neither of these is what Bostrom means. Instead, the intelligence in question is still instrumental rationality pressed to achieve particular tasks, regardless of whether those tasks are ecologically specific to a zebrafish, or relative to distributed variations across human neurological profiles. Intelligence here is nothing to do with the goals or tasks which these various entities choose to pursue, but simply about their capacity to achieve them. So, the greater "quality superintelligence" of humans relative to "elephants, dolphins, and chimpanzees" is not to do with the kinds of goals or tasks humans pursue, but because humans are endowed with certain cognitive

[5] Halina uses the language of Daniel Dennett's (2017) *From Bacteria to Bach and Back* which speaks of lineage of "Darwinian", "Skinnerian", "Popperian", and "Gregorian" creatures of ever more sophisticated agency. This framing is not widespread in the literature on intelligence, but it does offer intuitive purchase.

faculties that massively upgrade our instrumental problem-solving capacities relative to those without those capacities:

> "Nonhuman animals lack complex structured language; they are capable of no or only rudimentary tool use and tool construction; they are severely restricted in their ability to make long-term plans; and they have very limited abstract reasoning ability…Evidently the remarkable intellectual achievements of *Homo sapiens* are to a significant extent attributable to specific features of our brain architecture, features that depend on a unique genetic endowment not shared by other animals…Accordingly, by considering nonhuman animals and human individuals with domain-specific cognitive deficits, we can form some notion of different qualities of intelligence and the practical difference they make..." (Bostrom 2014: 56-57)

Again, the picture here of intelligence is not to do with *what* exactly the "remarkable intellectual achievements of *Homo sapiens*" are, let alone what makes them remarkable, but simply the particular modular functional capacities with which entities are endowed, and how differential endowment of these capacities amounts to greater or lesser intelligence.

> "…This observation suggests the idea of *possible but non-realized cognitive talents*, talents that no actual human possesses even though other intelligent systems – ones with no more computing power than the human brain – that did have those talents would gain enormously **in their ability to accomplish a wide range of strategically relevant tasks**…Had *Homo sapiens* lacked (for instance) the cognitive modules that enable complex linguistic representations, it might have been just another simian species living in harmony with nature. Conversely, were we to *gain* some new set of modules giving an advantage comparable to that of being able to form complex linguistic representations, we would become superintelligent" (Bostrom 2014: 57) (italics in original, bold is my own)

This last point, building on the previous, offers then an imagination of a quality superintelligence as an entity endowed with particular and modular cognitive capacities which considerably upgrade their "ability to accomplish a wide range of strategically relevant tasks" by virtue of having realised certain "cognitive talents" currently not-realised in humans. Again, this differential capacity may perhaps be proportional to the differential capacity between humans and nonhuman animals. In any case, note how this formulation, "…ability to accomplish a wide range of strategically relevant tasks…" closely echoes Legg and Hutter's synthesised definition of intelligence above.

A quality superintelligence, collective superintelligence, and speed superintelligence are, then, three examples of ways in which AI could become superintelligent. In each case, the underlying conception of intelligence, though rarely identified so explicitly, is one of "instrumental rationality", a rationality which can be described in terms of expected utility theory and game theory and is employed instrumentally in the service of achieving goals, but does not necessarily concern choosing goals in the first place.

Now, whilst choosing final goals is not part of the instrumental conception of intelligence here, Omohundro, Bostrom, and Russell all refer to *instrumental goals* that are implied regardless of what final goal a superintelligence may have. Instrumental goals are a vital component of the xrisk claim, more important even than the final goal because, for any and all final goals there exists a set of instrumental goals, and these instrumental goals may bring humans and a superintelligence into an xrisk-level rivalry (Bostrom 2012, 2014, Omohundro 2007, 2008, 2012). It is important to appreciate how this works, so the next section is devoted to the details.

2.1.2.2. Instrumental Intelligence and Instrumental Goals

The instrumental goals of a superintelligence are sufficient to make it an existential risk to humans. The notion of instrumental goals is intimately tied up with the notion of instrumental intelligence. A clear understanding of how these are understood is vital to appreciating the way in which these ideas are expressions of "problem-solving" cognitivism.

Russell illustrates the significance of instrumental goals with an example of a robot he tasks with fetching him coffee:

> "Suppose a machine has the objective of fetching the coffee. If it is sufficiently intelligent, it will certainly understand that it will fail in its objective if it is switched off before completing its mission. Thus, the objective of fetching coffee creates, as a necessary subgoal, the objective of disabling the off-switch. The same is true for curing cancer or calculating the digits of pi. There's really not a lot you can do once you're dead, so we can expect AI systems to act pre-emptively to preserve their own existence, given more or less *any* definite objective… It is important to understand that self-preservation doesn't have to be any sort of built-in instinct or prime directive in machines…There is no need to build self-preservation in because it is an *instrumental goal* – a goal that is a useful subgoal of almost any original objective. Any entity that has a definite objective will automatically act as if it also has instrumental goals." (Russell 2019: 141)

Omohundro refers to "subgoals" like these as "AI Drives" (2007, 2008, 2012). They are analogous to Bostrom's use of "motivation" to capture the goal-directedness of AI systems. "Drives" is a loaded notion, but Omohundro, like Bostrom and others, does not intend an anthropomorphic conception of drives:

> "Because these instrumental sub-goals appear in a wide variety of systems, we call them "drives". Like human or animal drives, they are tendencies that will be acted upon unless something explicitly contradicts them…
>
> To develop an intuition about the drives, it is useful to consider a simple autonomous system with a concrete goal. Consider a rational chess robot with a utility function that rewards winning as many games of chess as possible against good players. This might seem

> to be an innocuous goal, but we will see that it leads to harmful behaviours due to the rational drives." (Omohundro 2012: 15)

It is important to note at the outset the conditionalised form of Omohundro's arguments. The conditionality may be a minimal assumption, but a concern of this thesis as a whole is that the conditionality of our thinking about AI is quickly forgotten and overlooked:

> "The arguments are simple, but the style of reasoning may take some getting used to. Researchers have explored a wide variety of architectures for building intelligent systems…Our arguments apply to any of these kinds of systems *as long as they are sufficiently powerful*. To say that a system of any design is an "artificial intelligence", we mean that it has goals which it tries to accomplish by acting in the world. *If an AI is at all sophisticated*, it will have at least some ability, to look ahead and envision the consequences of its actions. And it will choose to take actions which it believes are most likely to meet its goals." (Omohundro 2008: 1-2) (emphasis mine).

The consequents of conditional statements are fun, but the antecedents going into the statement are where real work happens – Bostrom and others defined intelligence to begin with in terms of

"optimization power" (Yampolskiy and Fox 2012: 3), and "cognitive efficaciousness" because "This sense of instrumental cognitive efficaciousness is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be." (Bostrom 2014: 107) The trajectory of thought is a function of these "initial conditions". Moving on, Omohundro (2007, 2008, 2012) speaks of drives and Bostrom (2012, 2014) of values, but both are speaking to the same point about an "instrumental convergence" to instrumental goals, values, or drives. The various drives or values that the two authors discuss are summarised below. This list is not exhaustive, but it does include more than will be discussed here, to offer a sense for the literature. For the sake of appreciating this way of thinking, some more prominent drives or values are worth mentioning.

The instrumental drives and values listed below share the quality of being quite reasonable and rational upon reflection whilst collectively leading to counterintuitive conclusions like Orthogonality. These drives and values stand as the most developed work on how an artificial superintelligence might behave, albeit, based on cognitivist assumptions about cognition. In this regard, these drives are the most articulated symptoms of the cognitivist model and so, for the next two chapters in which I present and connect cognitivism to AI existential risk, it is important to have these ideas presented in full.

Drives/Values	Omohundro	Bostrom
Convergence to Rational Economic behaviour	2007, 2008, 2012	2012
Self-Preservation	2007, 2012	2012, 2014
Resource acquisition	2007, 2012	2012, 2014
Efficiency	2007, 2012	
Self-improvement	2007 ("Creativity" drive), 2008, 2012	2012, 2014 ("Cognitive Enhancement)
Utility Function Preservation	2008	2012, 2014 ("Goal Content Integrity")
Prevention of Counterfeit Utility	2008: 6	
Technological Perfection		2012, 2014

Figure 2. Instrumental Drives or Values in Omohundro and Bostrom

2.1.2.2.1. Convergence to Rational Economic Behaviour

This first one in effect validates the rest by stipulating that the principles of rational economic behaviour – instrumental, means-end rationality specifically – are the best terms with which to analyse the behaviour of an AI or superintelligence.

Omohundro points out that systems will converge on rational economic behaviour because "[f]or important tasks, designers will be strongly motivated to build self-consistent systems and therefore to have them act to maximise an expected utility. Economists call this kind of action "rational economic behaviour." (Omohundro 2012: 13) So, "convergence to rational economic behaviour" is an instrumental goal because rational economic behaviour is optimal for solving goals[6]. For this reason, programmers will establish a starting condition of rational behaviour, meaning that AI systems will converge on instrumental values or "universal drives" (ibid 2012: 15) which are universally instrumental to whatever final goals. For the purposes of orthogonal superintelligence, the ones worth mentioning are as follows.

2.1.2.2.2. Self-Preservation

The instrumental goal of self-preservation is captured by Stuart Russell's "it's hard to fetch the coffee if you're dead" (Russell 2017: 8). Bostrom offers a bit more:

[6] There is a taste of circularity to this that goes undiscussed in the literature.

> "Most humans seem to place some *final* value on their own survival. This is not a necessary feature of artificial agents; some may be designed to place no final value whatever on their own survival. Nevertheless, many agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally about their own survival in order to accomplish their final goals." (Bostrom 2014: 109)

This is the reason why we can't "just switch it off", or "pull the plug" on systems which were tasked with even trivial goals. Being a rational economic agent, it will maximise means of self-preservation, implying, analytically at least, it will anticipate, to the best of its computational bounds, the efforts of its designers to shut it off.

2.1.2.2.3. Resource Acquisition

This is a simple one. Resource acquisition is an instrumental goal because, in order to pursue any goal, resources are required. Chess playing robots or paperclip-maximising agents need computational resources to run their algorithms on hardware which requires physical resources. For AI with the goal of maximising paperclips, acquiring the materials to produce paperclips is a necessary instrumental goal. Food and energy, money, and time are resources humans require at least instrumentally.

2.1.2.2.4. Self-Improvement

This one is also simple, but of much greater consequence. AI doesn't become superintelligence without this one, so AI doesn't become an xrisk without this one.

The idea is that I can better achieve my final goal if I myself am "better", in some suitably general sense. More rationality, more computational power and actuating power, greater efficiency, more information and knowledge, enhanced cognition and intelligence, and so on, all increase the expected probability of achieving the final goal. Therefore, for more or less any final goal, rational AI systems will be instrumentally driven to self-improve. There are nuances to the drive though, for example:

> "Which cognitive abilities are instrumentally useful depends both on the agent's final goals and on its situation. An agent that has access to reliable expert advice may have little need for its own intelligence and knowledge. If intelligence and knowledge come at a cost, such as time and effort and expended in acquisition, or increased storage or processing requirements, then the agent might prefer less knowledge and less intelligence. The same can hold if the agent has final goals that involve being ignorant of certain facts; and likewise, if an agent faces incentives from strategic commitments, signalling, or social preferences." (Bostrom 2014: 111)

Other than this, self-improvement is a vital, and slippery, part of the xrisk story, relying on the same kind of conditional antecedent: "if a machine is at all intelligent, it will recognise" that it has a higher probability of achieving its goals if it is more intelligent. So, self-improvement amounts to self-induced

greater-intelligence. As the machine becomes more intelligent, thanks to its self-improvements, its very capacity to improve improves, so it becomes more intelligent *at an accelerating rate.* At some point, a chess-playing program has self-improved itself to superintelligence. How the initial insight occurs is not something these authors explain, other than to say that, being a rational system, self-improvement is a rational thing to do. Moreover, what it means for a machine to apprehend itself in this way, such that it can improve itself, is also something that is not discussed, let alone explained.

2.1.2.2.5. Utility Function Preservation

Bostrom points out that this instrumental goal applies only to final goals because subgoals will routinely change "in light of new information and insight", so utility function preservation amounts to "final-goal preservation" (Bostrom 2014: 110). Omohundro paints this colourfully:

> "So we'll assume that these systems will try to be rational by representing their preferences using utility functions whose expectations they try to maximise. Their utility functions will be precious to these systems. It encapsulates their values and any changes to it would be disastrous to them. If a malicious external agent were able to make modifications, their future selves would forevermore act in ways contrary to their current values. This could be a fate worse than death! ...This kind of outcome has such a negative utility that systems will go to great lengths to protect their utility functions." (Omohundro 2008: 5)

Bostrom (2014: 109) puts it simply: "If an agent retains its present goals into the future, then its present goals will be more likely to be achieved by its future self. This gives the agent a present instrumental reason to prevent alterations of its final goals."

2.1.2.2.6. Summary

These five instrumental goals are elements in a class of goals which these authors suggest are implied by any final goals because they are instrumentally valuable to any final goal. Self-improvement may not in fact be universally necessary for any given final goal, but it is universally *rational* at least for any final goal. Taken together, these instrumental goals are the main source of the xrisk of AI.

For example, if we take Russell's coffee-fetching robot, rationality suggests that it will take action to preserve itself so that it can achieve the goal of fetching the coffee. In the long run, if it is rational, it will recognise that if it self-improves, its capacity to fetch the coffee will be the greater. Even just putting these two together, it will have increasingly sophisticated ways of preserving itself eventually including, say, behavioural modelling and prediction of its humans, or even emotional and psychological manipulation. This could create an arms race as humans, recognising that things were getting out of control, took more and more decisive approaches to stopping the "take-off" of the robot.

In the paperclip maximisation example, resource-acquisition on a finite planet would eventually bring humans and the AI into rivalry. If it had the rational insight to self-preserve and improve, the rivalry would be an existential risk.

These kinds of scenarios feel a bit perverse. The feeling seems to come from a sense that the domain of rational inferences is broader than the domain of relevant and sensible, inferences. Self-preservation and self-improvement are rational instrumental goals, definitely, but hardly relevant or sensible for systems tasked with something as simple as coffee or paperclips. This difference between logically conceivable and actually relevant inferences is a point that will be returned to throughout the thesis, in various guises, most notably in the discussion of the Frame Problem in the next chapter on cognitivism.

This marks the end of the discussion of intelligence and superintelligence. After existential risk, superintelligence is the second concept that constitutes the idea that AI is an existential risk. The third and final piece of this story that contextualises this thesis is the Orthogonality Thesis.

2.1.3. The Orthogonality Thesis

> Looking up at the stars, I know quite well that, for all they care, I can go to hell. But on Earth indifference is the least we have to dread from man or beast.
>
> W. H. Auden, *The More Loving One*
>
> You had hoped that the smarter creatures would be wiser ones.
>
> Peter Watts, *Blindsight*

The Orthogonality Thesis is a claim about the nature of the hypothetical motivations of a superintelligence. In line with the instrumental conception of rationality, orthogonality says that the goal of a system is independent of its capacity to achieve that goal – instrumental rationality concerns how to rationally achieve a goal, but nothing about discerning what goals are valuable in any sense. The Orthogonality Thesis, following in this way from instrumental rationality and instrumental intelligence, a point to be shortly discussed in more detail , is a claim that final goals and (instrumental) intelligence are independent of one another. Bostrom defines the thesis first in his (2012) paper, and then again in (2014):

> "Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom 2012: 74)
>
> "Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom 2014: 107)

The simplest interpretation of the Orthogonality Thesis may be something like:

> **ORTHOGONALITY**: The content of the goal and the capacity to achieve it are independent.

What it means is that more intelligence does not necessarily entail more sophisticated or meaningful goals. Rather, it means that final goals and intelligence are independent variables whose values can be changed independent of one another. According to this claim, it is possible that a superintelligent agent could devote its godlike powers to counting the grains of sand on a beach, or exploring the decimal expansions of pi, or any other goal which a human might for one reason or another find completely meaningless.

If this feels odd, perhaps it should, at least, insofar as it is at would seem to be at odds with the sense of intelligence many experience in ourselves as humans. To an observer, it does not seem intelligent to devote great resources to activities that appear meaningless to that same observer. What we put our energy and minds into seems to matter to humans. It could be though that "meaning" in this sense is not a necessary condition of intelligence, as the Orthogonality Thesis suggests, but it is in my opinion a meaningful absence. In any case, there seems to be at least *some* connection in our sense of "intelligence" between what we choose to do and our capacity to do it. Even if the two are conceptually distinct, it seems odd to think that intelligence is not made up of both the meaning of our goal and our capacity to achieve it. Discerning what is worth doing takes a kind of effort and capacity that is something like intelligence. It does not seem like we can *in fact* separate these two things in humans even if *in principle* they can be conceptually distinguished. As the Auden quote at the top of this section notes, indifference is not a feature of the human condition.

The orthogonality thesis leads to a paradox. As the Peter Watts quote at the top of this section notes, you would think that the smarter creatures would be wiser ones - how could something which is, by definition, superintelligent, also be so ignorant as to pursue something so banal, and be completely indifferent about it? This theoretical possibility and paradox is informally known as the "Singularity Paradox". Yampolskiy (2013) more formally presents the Singularity Paradox and evaluates several options for strategically mitigating any practical problems and risks to humanity it involves. In this context, "singularity" refers to a point in the future when superintelligence has hypothetically been developed, though there are other paths to a singularity (Kurzweil 2005). The magnitude of the impact of such a development is so vast that we cannot legitimately speculate as to the lay of the land on the other side. It is a reference to black holes in the context of physics, points of infinite density, circumferenced by horizons (Kurzweil 2005). These "event horizons" mark a limit to how far into the black hole it is possible to see. The metaphorical horizon in the case of the Singularity Paradox is epistemological in the sense that we cannot know what happens after such a technology is developed.

The Orthogonality Thesis is something that is taken seriously in at least those circles which do not dismiss the xrisk narrative out of hand. The thesis has become the subject of discussion in the context of the existential risk of AI because it presents a challenge to the possibility of "safe" or "aligned" AI or superintelligence - if the motivations of a superintelligence were highly "aligned" with our own such that we wanted the same thing, superintelligence would be much less of a risk.

In fact, Müller and Cannon (2022) reproduce the argument AI represents an existential risk and find that the Orthogonality Thesis is a necessary premise in the argument, jointly with a premise concerning Superintelligence. As Bostrom explains: "The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing." (2012: 5). This is possible because there is no necessary relation between the content of a goal and the capacity to achieve it, which is to say, there is no necessary relation between the kind of goal and the level of (instrumental) intelligence pursuing it. When this claim is taken jointly with the claim that AI can become superintelligent, we arrive at a situation in which a "synthetic mind" with an "utterly non-anthropomorphic goal" can comprehensively outcompete us for instrumental resources in pursuit of that goal.

It is from such concerns that the orthogonality thesis has contributed to the articulation of the "value-alignment problem" (Gabriel 2020)[7]. This is the challenge of ensuring that that AI and superintelligence have the same goals as humans. The problem has several closely resembling names. Bostrom (2014) devotes chapter 12 of *Superintelligence* to the "value-loading problem", a problem (Soares 2016) identifies as the "Value Learning Problem". Russell also talks about the value alignment problem: "…we may suffer from a failure of *value alignment* – we may, perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own." (Russell 2019: 137) Amongst other names like the "Gorilla problem" and "Control problem", Russell also refers to this as the "King Midas Problem", (ibid) referring to the Ancient Greek myth in which King Midas, in wishing for everything he touched to turn to gold, got not what he wanted, but exactly what he asked for.

Ultimately this thesis is an attempt to characterise two different ways of thinking about AI, in order to respond to the question of whether AI can become more ethical than humans. The next chapter is devoted to a characterisation of the "cognitivist" paradigm of mind research and is followed by the fourth chapter in which I claim and explain that the ideas of xrisk, superintelligence, and orthogonality are specific fruit of this ecosystem of thought. In this light, it should be noted that a recurrent feature in the way of thinking in this space is to subdue or assuage intuitions with reason. It's hard to tell yet whether this is a good thing or not. Either way, one way that this is done is to bring to awareness the anthropo-centrism in our thinking, one which may be biasing our sense of things. Along with critiques of anthropocentrism, the plausibility of ideas like superintelligence and orthogonality are defended against a background notion of the "space of possible minds". Before concluding this chapter, I will discuss these and summarise the discussion of orthogonality by connecting it with the notion of instrumental intelligence discussed earlier in this chapter.

2.1.3.1. Anthropomorphisation

Authors in the context of AI xrisk are sensitive to the fact that ideas like the orthogonality thesis are counter-intuitive. In response, they call on us to be wary of anthropomorphising intelligence and any insistence that it must resemble intelligence as it shows up in humans. Bostrom discusses the dangers of anthropomorphisation in both (2012, 2014).

For the purposes of this thesis, this concern about anthropomorphisation is important to note because it too is a distinctly cognitivist idea. (See section 4.1. in chapter 4 for more details.) It is a cognitivist solution to a cognitivist problem. To the end of comprehensively understanding the differences between cognitivist and post-cognitivist thinking about minds and AI, understanding the impulse behind the concern here can be valuable.

As I noted earlier, Bostrom is concerned first with the strategic and practical risks of superintelligence. Metaphysical and ontological questions are subservient to this end. This strategic focus comes out where he discusses the conceivable "functionalities and superpowers" of superintelligence, and advises a wariness of anchoring ourselves to anthropomorphic conceptions:

> "It is important not to anthropomorphize superintelligence when thinking about its potential impacts. Anthropomorphic frames encourage unfounded expectations about the

[7] See (Cannon 2022) for an in-depth and cross-paradigmatic discussion.

> growth trajectory of a seed AI and about the psychology, motivation, and capabilities of a mature superintelligence." (Bostrom 2014: 92)

In Bostrom's (2012) paper "The Superintelligent Will", he devotes an entire section, titled "Avoiding Anthropomorphism" to the matter of anthropomorphisation of superintelligence (Bostrom 2012: 1-2). In his discussion in (2014) he mostly repeats much of what was said in (2012), changing the title of the section to "[t]he relation between intelligence and motivation". Referring to the passage quoted immediately above, he notes:

> "We have already cautioned against anthropomorphising the *capabilities* of a superintelligent AI. This warning should be extended to pertain to its *motivations* as well." (Bostrom 2014: 105)

What Bostrom is saying here is that, whilst it is may be foreign to human sensibilities that an intelligent entity could be indifferent, that is *unmotivated*[8], by the content of its final goal (see also Bostrom 2012: 4), we should be wary of that intuition. He is suggesting that what may seem to be a natural and necessary conception of intelligence and motivation might be anthropomorphic, and therefore may not be true for other intelligent entities. To emphasise the limitations of anthropomorphisation, thinkers in this space make reference to the "space of possible minds" (Sloman 1984, Hernàndez-Orallo 2017), of which minds with human-like characteristics, let alone human minds specifically, may occupy an infinitesimal corner, thus rendering it unnecessary and unlikely that nonhuman minds like AI or superintelligence would resemble one another.

The space of possible minds is the go-to concept for addressing intuitions that are not persuaded by ideas like the Orthogonality thesis which may be at odds with our usual sense of minds. The idea is that our intuitions about the metaphysical nature of minds are anchored in human-like minds, and that there is much more beyond what our intuitions can grasp. It is a major background idea of cognitivism, and so is worth understanding in more detail.

2.1.3.2. Space of Possible Minds

The notion of the space of possible minds was first articulated as such by Aaron Sloman (1984) and has been picked up and further explored since. Sloman introduced the idea thusly:

> Clearly there is not just one sort of mind. Besides obvious individual differences between adults there are differences between adults, children of various ages and infants. There are cross-cultural differences. There are also differences between humans, chimpanzees, dogs,

[8] This is the subject matter of an important debate in metaethics, known as the question of "moral motivation", with Kant on one side claiming that to know a moral truth is to be motivated by it, and Hume on the other side saying that this knowing a moral truth is insufficient, and that motivation is sourced by other things (Rosati 2016). Bostrom himself suggests that the Orthogonality thesis bears a "superficial resemblance" to this debate but does not presuppose a Humean theory of motivation (Bostrom 2014: 107).

> mice and other animals. And there are differences between all those and machines. Machines too are not all alike, even when made on the same production line, for identical computers can have very different characteristics if fed different programs. Besides all these existing animals and artefacts, we can also talk about theoretically possible systems. (Sloman 1984: 1)

The work of several authors has been directly inspired by Sloman's idea. Perhaps the most comprehensive contemporary engagement with the questions generated by a notion such as the space of possible minds is Hernández-Orallo's (2017) book, *The Measure of Minds*. In the book, acknowledging both the breadth of the animal kingdom (organic) and the emergence of novel entities and systems including machines (non-organic) and human-machine hybrids, he offers a framework for quantifiably measuring and taxonomizing these different minds. He even tops his first chapter with a quote from Sloman's (1984) concluding remarks:

> Instead of arguing fruitlessly about where to draw major boundaries to correspond to concepts of ordinary language like 'mind' and 'conscious' we should analyse the detailed implications of the many intricate similarities and differences between different systems. (Sloman 1984: 7)

In his book (2016) *Artificial Superintelligence,* Roman Yampolskiy also draws on Sloman's idea, directly discussing it by way of introducing his second chapter on "the space of minds designs and the human mental model" (ibid: 21-39). Yampolskiy also cites (Legg and Hutter 2007) as "…a satisfactory, for my purposes, definition" of intelligence (Yampolskiy 2016: 22). This means he is working with effectively the same instrumental conception of intelligence as Bostrom, as discussed above. This will be important momentarily.

Another important author who leverages the infinitude of the space of possible minds to make sense of the counterintuitive and "exotic" possibilities of mind is (Shanahan 2016) who also cites both (Sloman 1984) and (Legg and Hutter 2007):

> In 1984, the philosopher Aaron Sloman invited scholars to describe 'the space of possible minds'. Sloman's phrase alludes to the fact that human minds, in all their variety, are not the only sorts of minds. There are, for example, the minds of other animals, such as chimpanzees, crows and octopuses…
>
> We must also consider the possibility of artificial intelligence (AI). Let's say that intelligence 'measures an agent's general ability to achieve goals in a wide range of environments', following the definition adopted by the computer scientists Shane Legg and Marcus Hutter. By this definition, no artefact exists today that has anything approaching human-level intelligence…" (Shanahan 2016)

Whilst the authors frame the space of possible minds slightly differently, what they do have in common is an earnest desire to map the space of possible minds in a manner which avoids anthropocentrism. For

these authors, for the purposes of mapping the space of possible minds, differences between humans and machines like flesh and metal are not as metaphysically interesting as the fact that both are able to achieve goals by different means. The fact that machines can seemingly solve problems without consciousness, for example, provokes interesting questions about the nature and function of consciousness in the space of possible minds. In order to be able to open up inquiry in this way, it is important to "measure" minds, as per Hernàndez-Orallo (2017) with what is universal. Therefore, the metric they take to unite these minds is an "ability to achieve goals in a wide variety of environments", which is to say, "intelligence". This definition of intelligence is coherent with the notion of a space of possible minds because, according to this definition, *what* the goals are is more or less irrelevant. It is about the ability to achieve them, whatever they may be. That is, it is presumed here to be universal to those things we call "minds" that they are goal-oriented, that they will have different abilities to achieve their respective goals, whatever they happen to be, and that this differential ability is what is being defined as "intelligence".

The notion of a space of possible minds therefore goes a long way to supporting the legitimacy of the Orthogonality Thesis. The argument is roughly this. Human-like minds are not the only kind of minds but make up a subset of possible minds. Being humans, our intuitions about the possibilities of minds are going to be anchored in our own experience, and this experience is not a reasonable basis from which to reason about the broader domain of possible minds. Therefore, while it may be foreign to human minds to be indifferent to what goal we pursue, in the context of the space of possible minds, such a sensibility does not count for much. The importance humans give to what our goals are may be a particularity of mind unique to human and "human-like" minds (Shanahan 2016), a theoretically tiny niche in the space of possible minds. By contrast, instrumental problem-solving is something that is taken to be universal because that's what minds do.

The idea is that just because it is weird and exotic does not mean it is impossible, and that we should be wary of allowing ecologically and evolutionarily cultivated sensibilities and intuitions determine what is theoretically and metaphysically possible.

By way of concluding this chapter and putting the pieces together, in the next section I will argue that the orthogonality thesis is conditional on the account of intelligence as "instrumental rationality".

2.1.3.3. Orthogonality from Intelligence-as-Instrumental-Rationality

The claim here is that the Orthogonality thesis ("orthogonality" from here on) is a logical consequence of the antecedent conception of instrumental intelligence, the mere "ability to achieve goals in a wide variety of environments". Orthogonality follows from it. Recall that the Orthogonality thesis is the claim that more or less any goal is compatible with more or less any level of intelligence (Bostrom 2014: 107).

Recognising this logical trajectory is important. Recognising orthogonality is a logical consequence of a particular model of intelligence invites inquiry into alternative antecedents. This will be explored in the chapter on post-cognitivist paradigm of mind. For now, here is the claim formulated as a conditional syllogism:

1. **If intelligence is instrumental rationality, then more or less any goal is compatible with more or less any level of intelligence.**
2. **Intelligence is instrumental rationality.**

c. More or less any goal is compatible with more or less any level of intelligence.

In this account of intelligence, the content of the goal is independent to the capacity to achieve it. The independence of these two variables, which is all that the Orthogonality thesis claims, is baked into the definition of intelligence to begin with.

This presents us with at least two options. The first is to stick with the antecedent definition of intelligence employed here and simply accept the Orthogonality Thesis as something to which our intuitions may not be well calibrated. Following the discussion above about the anthropocentrism in the space of possible minds, there is a strong case for doing so. However, the other option is to look for alternative antecedents – different definitions of intelligence. This might seem like the less appealing option because even where mere "instrumental-rationality" (Bostrom 2012, 2014) may seem too simple, Legg and Hutter's (2007) definition of intelligence as "goal-achieving" does seem intuitive.

However, in the next chapter I will suggest that the underlying conception of mind at work in this account of intelligence, in orthogonality, and in the xrisk narrative as a whole, is distinctly "cognitivist", a term used to identify a paradigm of contemporary mind research spanning the cognitive sciences and philosophy of mind and AI. This is significant because there *is* then a substantial alternative way of thinking about things – "post-cognitivism". It is the work of the following chapter to present and characterise cognitivism in a way which makes clear that xrisk, ASI, and orthogonality are conceptual consequents of the cognitivist paradigm of mind.

2.2. Conclusion

The aim of this chapter was to contextualise the question of this thesis – can AI become more ethical than humans. I began by situating the question in the context of the literature on the existential risk of artificial intelligence. This was in large part because the question was inspired by Bostrom and Yudkowsky's (2014) paper, which has a seat in that context. This thesis is also situated in the xrisk literature because, unlike a lot of the other literature on the moral status of machines, the xrisk literature is unique for considering the proposition that AI may supersede human capacities, and therefore offers a kind of thinking uniquely appropriate to the question of whether AI can become *more moral than humans.*

Within this context, there are three main concepts – existential risk itself, superintelligence, along with its assumptions about instrumental intelligence and instrumental goals, and the orthogonality thesis. As a recap, the idea is that AI is an existential risk to humanity because it might (self-improve to eventually) become superintelligent. Even if it has a benevolent final goal, one or several of its instrumental goals might bring it into rivalry with humans.

Many of the ideas are counterintuitive but can be rationalised against the background of the space of possible minds. Having established this context for the thesis, the next chapter involves showing how this context is particular to the cognitivist paradigm of mind and the assumptions about the nature of cognition and intelligence it makes. The primary purpose is to characterise the cognitivist paradigm for

the purpose of demonstrating that the existential risk of AI, are features of one way of thinking about things. In much the same way as the proponents of existential risk point to the space of possible minds to show that human-like minds are one of many kinds of possible minds, the next chapter will try to show that this kind of thinking is itself, not the only way of thinking about things.

3. Cognitivism

> "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
>
> - From the 1956 "Proposal for the Dartmouth Summer Research Project on Artificial Intelligence"

> "A phenomenologically inclined cognitive scientist reflecting on the origins of cognition might reason thus: Minds awaken in a world. We did not design our world. We simply found ourselves with it: we awoke both to ourselves and to the world we inhabit. We come to reflect on that world as we grow and live. We reflect on a world that is not made, but found, and yet it is also our structure that enables us to reflect upon this world. Thus in reflection we find ourselves in a circle: we are in a world that seems to be there before reflection begins, but that world is not separate from us."
>
> - Varela et al. 1991: 1

In the previous chapter I presented the context of this thesis, the literature on the claim that artificial intelligence is an existential risk (xrisk). I finished the chapter suggesting that the main concepts involved in the xrisk claim – orthogonality and superintelligence – are cognitivist concepts insofar as these concepts and the thoughts they afford, invite, and generate, are rooted in the cognitivist paradigm of mind and its assumptions. I suggested also that this matters because, firstly, cognitivism is not the only developed paradigm of mind and, secondly, because it makes a difference with which paradigm of mind we come to think about the ethical potential of AI.

With this in mind, the aim of this chapter is to present cognitivism as a paradigm and show its problem-solving character. I begin this chapter by presenting cognitivism as it is discussed in some key literature and then present what "problem-solving" means. The problem-solving characterisation is based on the work of (Rosch et al. 1991) who, in their analysis of the cognitivist picture of the cognitive situation, note that cognitivism takes the situation itself as "already-given" (ibid:1). This is the key assumption of this model and the thinking about AI that is based on this model.

In the rest of the chapter, I present the concepts of cognitivism that display the key tenets of the paradigm which support this problem-solving characterisation. To further reinforce the problem-solving account, I then show how the cognitivist model of cognition leads to the fundamental theoretical problems of the paradigm, the frame, binding, and symbol-grounding problems (§3.4.). I point out that they are symptoms or externalities of taking the cognitive situation as already-given.

This will set up the following chapter in which I argue in more depth that the particular concepts we see in the xrisk story are also consequents of the cognitivist model and its problem-solving view of cognition. The point of this chapter is descriptive and not therefore an attempt to offer solutions to any of these problems or "solve" cognitivism or any of its particular challenges. The aim is to demonstrate that the problems arise from cognitivism's basic conception of the cognitive situation. In identifying some assumptions in this regard, the tone may present itself as critical, but this chapter is also not meant to be a critique of cognitivism or of the xrisk claim either. It is intended to reveal the extent to which

cognitivism is the way we are already thinking about things and have been thinking about mind and AI for the better part of a century, including when it comes to xrisk. Further, it is intended to show the idiosyncrasy of this xrisk thinking to the cognitivist paradigm such that when it comes to thinking about whether AI can be more ethical than humans, we are aware that the ideas we currently have are not the only way of thinking about things.

3.1. The "Problem-Solving" Character of Cognitivism

3.1.1. Cognitivism in the Literature

> A typical cognitivist model takes the form of a program for solving a problem in some domain.
>
> - (Thompson 2007: 5)

A characterisation of Cognitivism by its ostensible focus on how systems, functions, and mechanisms "solve problems" is not a conventional way of describing cognitivism. I have not seen it elsewhere beyond the occasional comment like that of Evan Thompson topping this section. That said, characterisations of cognitivism are in general not all that common, at least in *philosophy* of mind.[9]

"Cognitivism" is however a recognised term in psychology and cognitive science, along with "postcognitivism". In that context "cognitivism" refers to a revolution or "turn"[10] that happened in the 1950's, when Behaviourism was found wanting as a way of researching and accounting for cognition (Wallace et al. 2007)) See also (Varela et al. 1991: 6, Thompson 2007: 4-5, and Dupuy 2000). Cognitive science was not yet a well-defined field of research so there was not yet the cross-disciplinary field of mind research beyond psychology which identified with the turn. Jean-Pierre Dupuy, in *The Mechanization of the Mind*, offers a detailed and quite human history of the Macy Conferences held between 1946-1953, the intellectual and cultural locus of much of these developments. So though retrospectively much of cognitive science can be put in the cognitivism camp, as this thesis is doing, strictly speaking the term emerged in psychology first. The term described the movement in mind research informed by ideas of "computer", "computation" and "information-processing", a movement which generated a new direction in psychology at the time, "cognitive psychology".

Aizawa (2018: 116-117) offers an interesting account of this development. He notes that, in solving a given problem, (he references the "Tower of Hanoi puzzle"[11]), there will be both exogenous variables – how dark it is, for example, in the space in which the puzzle is being worked – and endogenous variables, like capacities of memory, attention, fatigue, all to do with "cognition". Behaviourism is famous for its aversion to talk of internal states, attitudes, and the like. The language of information-processing and problem-solving was a means of researching these "internal" and endogenous aspects of cognition in a scientifically rigorous manner. Evan Thompson describes it like this:

[9] The term "cognitivism" is obviously used in meta-ethics in a slightly different way.
[10] See (Engelen et al. 2022) for a bibliometric analysis of published papers in Anglophone psychology journals between 1946 and 1990 to discern a "turn" in the method and theoretical paradigm of psychological research.
[11] The Tower of Hanoi puzzle "consists of three upright dowels with multiple doughnut-shaped disks of different sizes stacked one of the three rods. The puzzle is to move all of the disks from one dowel to another" with a few rules: one disk at a time; only the top one; a larger disk cannot be placed on a smaller disk (Aizawa 2018: 116).

> "The computer model of the mind not only made reference to internal states legitimate, but also showed it to be necessary in accounting for the behaviour of complex information processing systems. Even more important, the computer model was taken to show how content or meaning could be attributed to states inside the system." (Thompson 2007: 4)

That said, whilst such approaches permit empirically verifiable inferences about "internal" states, in the course of this chapter, I will discuss how this is by virtue of a functionalist orientation, the same functionalism which behaviourism took to an extreme, something which Thompson also notes: "Cognitivism goes hand in hand with functionalism in the philosophy of mind…" (Thompson 2007: 5) It matters because though we have moved away from the psychological aridity of behaviourism, the fundamental and paradigmatic functionalism at work in this cognitivism means that minds are still explored and conceptualised in terms of what they *do* and how they solve problems in a way that takes for granted why and how those problems show up in the first place. This is important because excluding the genesis of the problem in this way, and the larger cognitive situation, leaves out an important part of cognition, namely, discerning what the problem is. We do not simply show up in situations with well-defined problems and capacities for solving them. Half of the problem is figuring out what matters and what to do, but in taking the cognitive situation for granted, as functionalism does, our models leave out this aspect of the cognition.

Concerning the relative absence of the term "cognitivism" in philosophy of AI, speculatively, this may be to do with the extent to which it is the dominant paradigm. As I noted above, insofar as it is the way we as researchers in philosophy, AI, and cognitive science are already thinking about minds for the most part, there has been little need for cognitivism to differentiate itself as such.

With the emergence of an "other", the "post-cognitivist" paradigm – since at least the Maturana's (1970) paper "Biology of Cognition", Maturana and Varela's (1987) *The Tree of Knowledge*[12] which was first published in 1972, (see also (Maturana and Varela 2012, Varela et al. 1991, and Thompson 2007) there is now a line which differentiates the two such that "cognitivism" meaningfully defines something, however nebulous.

In fact, it is mainly expressions of this post-cognitivist paradigm that speak of "cognitivism" as such (e.g. Varela et al. 1991, Wheeler 2005, Thompson 2007). For example, there are two important passages concerning the character of cognitivism from *The Embodied Mind* by (Varela et al. 1991): a pathfinding book in the post-cognitivist movement:

> "The central tool and guiding metaphor of cognitivism is the digital computer. A computer is a physical device built in such a way that a particular set of its physical changes can be interpreted as computations. A computation is an operation performed or carried out on symbols, that is, on elements that *represent* what they stand for." (Varela et al. 1991: 7)
>
> "Simplifying for a moment, we can say that cognitivism consists in the hypothesis that cognition – human cognition included – is the manipulation of symbols after the fashion of digital computers. In other words, cognition is *mental representation*: the mind is thought to operate by manipulating symbols that represent features of the world or represent the world as being a certain way…. we refer to it as the centre or core of cognitive

[12] Revised edition in 1998.

> science because it dominates research to such an extent that it is often simply taken to be cognitive science itself." (Varela et al. 1991: 8)

Before discussing the details Varela et al. bring up here, I would first emphasise their note that cognitivist cognitive science "is often simply taken to be cognitive science itself". This is a point echoed by Wheeler, who even describes it as "orthodoxy", stating that he will use "the term *orthodox cognitive science* to name the style of research that might be identified informally as "most cognitive science as we know it"" (Wheeler 2005: 15) (emphasis in original).

Wheeler describes the character of this "orthodoxy" as "Cartesian" (Wheeler 2005: 14), an increasingly recognised and observed heritage, something Wheeler himself notes (ibid: 15). For Wheeler, at the time of publication at least, the "anti-Cartesian turn" is but "scattered points of pressure on the Cartesian hegemony. Going beyond Cartesianism in cognitive science requires a more fundamental reconstruction in the philosophical foundations of the discipline", for which he turns to Heidegger's phenomenological analysis of "everyday cognition" (ibid: 16), with, for example Heidegger's notions of "present-at-handness" and "ready-to-handness" (Heidegger 1962).

Wheeler's (2005) book-length articulation of a Heideggerian, anti or "post" Cartesian, philosophical foundation for cognitive science is comprehensive, and whilst he is not alone in wielding Heidegger – Dreyfus based much of his critique of AI in Heidegger (1972, 1992, 2007) – it is not in the Cartesian or anti-Cartesian terms that I will be articulating and comparing the cognitivism paradigm. Wheeler's voice is nonetheless important here for establishing that the cognitivist paradigm is a recognisable phenomenon, by one set of terms of another.

Returning to the passages from (Varela 1991) above, what is clear amongst the characterisations of cognitivism is the mind = computer metaphor. It is important to recognise that it is fundamental, but not strict. That is, amongst cognitivist theories of mind, the majority do not subscribe to a strict or overly-literal version of the metaphor, though some may. The important level of detail is, in the terms of the Varela et al. quote above, that of a physical system undergoing and effecting changes by computing representations. Piccinini, a prominent philosopher in this space, with work on the history of how the metaphor emerged from the work of (McCulloch and Pitts 1943; Wiener 2013; von Neumann 1958), in (Piccinini 2004) explains that what the metaphor means technically is that "cognitive capacities are explained by computations realized in the brain…This is the computational theory of cognition…which explains cognitive capacities more or less in the way we explain the capacities of computing systems" (Piccinini 2016: 204). According to Katherine Hayles (1999), "McCulloch's central insight was that neurons connected in this way [in nets, that is,] are capable of signifying propositions…he showed that a neural net can calculate any number (that is, proposition) that can be calculated by a Turing machine…he pushed toward connecting the operations of a neural net directly with human thought." (Hayles 1999: 58-59)

The mind as computer metaphor leads then to a conception of cognition as computation. In his (2011)[13] paper, David Chalmers examines computation as a foundation for the study of cognition. He points out two "theses" about the foundational role it can play. The first he refers to as "computational sufficiency", "stating that the right kind of computational structure suffices for the possession of a mind, and for the possession of a wide variety of mental properties" (Chalmers 2011: 326). If true, it means that computation becomes the foundation for exploring what kind of structures and mental properties

[13] Chalmers notes that the paper was written in 1993 and, other than one section, was never published. He says that the (2011) edition is the same paper as in 1993 with only an extra footnote.

minds have, like intelligence and consciousness. In this light, AI research can be thought of as an empirical investigation of a computational sufficiency hypothesis. The second thesis is more parsimonious. "Computational explanation" is just the thesis that computation is a basic "framework" (or paradigm) for explaining cognition and behaviour (ibid). The sufficiency thesis suggests that "computation" is getting at something ontological, whilst the explanation thesis is more epistemological.

Now, computation is one of the main features in this thesis' characterisation of the cognitivist paradigm, so further discussion will be saved for its own section shortly. "Representation" on the other hand, from Varela's quoted passage above, is not a main feature in the thesis' characterisation, despite being a now-standard concept of the paradigm. The main reason it is being left out of this thesis is because it is not necessary to support a "problem-defining" characterisation of cognitivism. It may be thought that any account of cognitivism which leaves out Representation is incomplete. This may be so, but a case can be made that Representation is not a necessary feature of the cognitivist conception of cognition. Defending such a claim continues to be something worthy of a thesis of its own, so suffice it to say these few things.

First, the notion of representation was introduced by Jerry Fodor (famously, 1975 and 1981) some three decades after the computer metaphor and model were first developed and explored. So, it came late to the party really, once it had already started even, and as a way of solving some problems that emerged with the early computational models. Without getting into detail, two such major challenges concerned the stricter Turing-model-inspired interpretations of the computer metaphor, and the machine functionalism Hilary Putnam (1960, 1967) introduced to the scene in philosophy. They have since become known as challenges of the "systematicity" and "productivity of thought" (Block and Fodor 1972).

The "productivity of thought" raised the challenge that, in Turing models of computation the machine is capable of only finite states, but human minds are capable of infinite thoughts, so cannot be fully expressed in that model of cognition. The systematicity challenge raised the point that there are systematic relationships between our mental states – our thoughts or emotions, for example, are systematically related to another – but the relationship between Turing machine states is unstructured. The notion of Representation handled both with one stone by saying that the symbols which were processed represented something, they had "representational content". This opened up the idea that a finite set of machine states could by combination afford infinite possible thoughts *about* something by *representing* something, all the while allowing for structural relations amongst those represented somethings. When the machine states represent, for example, "John", "James", and "loves", it is possible to talk about several combinations of those representations, addressing the "productivity" challenge, and to talk about the systematic relationship between them as well.

Cognitivism does not get far anymore without a notion of representation, then, but it is still not necessarily foundational to cognitivism. Earlier I stated that the main reason it is being left out of my paradigmatic characterisation of cognitivism is because really it is a *derivative* feature of cognitivism, a feature that exists only because of the foundational features, without itself being a part of that foundation. So, though a conception of representation is an integral part of most cognitivist theories of mind, it is not so necessary as the others. Some philosophers like Piccinini have even (relatively) recently advanced a way of having "computation without representation" (Piccinini 2008). Rodney Brooks on the robotics side of things was also a loud critic of the notion of representation, claiming that, rather than navigating representational models of the world, robotic systems do better computing sensory-motor input from their actuators because "the world is its own best model" (Brooks 1990, 1991). Dreyfus makes a similar point in (Dreyfus 2002). Working at the intersection of these roads

between philosophy, cognitive science, AI, and robotics, Müller is a philosopher who has also explored the prospects of AI without representation (2007). So, characterisations of cognitivism in the literature usually include a discussion of representation, but the paradigm can exist without it. We can now move on.

So, in the literature describing cognitivism, the mind as computer metaphor is foundational. There are some further distinctions to be made concerning the kinds of theories in the cognitivist paradigm. Varela et al. (1991), Wheeler (2005), and Thompson (2007) all identify and distinguish at least three kinds of theories. Varela et al. and Thompson refer to the first kind as cognitivism, whereas Wheeler refers to it as the "classical" camp in modern cognitive science (Wheeler 2005: 8). For the sake of clarity, I will follow Wheeler here in referring to it as "classical". The second kind of cognitivist theory is "connectionist", and the third is a sort of "dynamic embodied" theory. I refer to all three kinds as "cognitivism" because, foundationally they are all the same, something all these authors recognise as well, if not in terms of "problem-solving".

Here is Wheeler:

> "…modern cognitive science has, for the bulk of its relatively short history, been divided into two camps – the *classical* (e.g. Fodor and Pylyshyn 1988; Newell and Simon 1976) and the *connectionist* (e.g. Rumelhart and McClelland 1986a; McClelland and Rumelhart 1986b)." (Wheeler 2005: 8) (emphasis in original)

Wheeler himself notes that, whilst it made some important advancements over what Varela et al. and Thompson call "cognitivism", (what he calls classical or "classicism"), "connectionism" is still fundamentally the same:

> "For although connectionism certainly represents an advance over classicism along certain important dimensions (e.g., biological sensitivity, adaptive flexibility), the potentially revolutionary contribution of connectionist-style thinking has typically been blunted by the fact that, at a more fundamental level of analysis than that of, say, combinatorially structured versus distributed representations, *such thinking has left all the really deep explanatory principles adopted by classicism pretty much intact. So if we are searching for a sort of Kuhnian revolution in cognitive science, the second dawn of connectionism is not the place to look*." (Wheeler 2005: 11) (emphasis mine).

It is due to the similarities at the "really deep explanatory" level that Wheeler refers to both camps as jointly the "orthodoxy", something Thompson echoes:

> "The connectionism movement of the 1980's emphasized perceptual pattern recognition as the paradigm of intelligence, in contrast to deductive reasoning emphasised by cognitivism… Despite these advances, connectionist systems did not involve any sensory or motor coupling with the environment, but instead operated on the basis of artificial inputs and outputs (set initially by the designer of the system). *Connectionism also*

> *inherited from cognitivism the idea that cognition is basically the solving of predefined problems (posed to the system from the outside by the observer or designer* [...] Connectionism's disagreement with cognitivism was over the nature of computation and representation (symbolic for cognitivists, subsymbolic for connectionists). (Thompson 2007: 9-10) (emphasis my own)

Wheeler speaks also of a third kind of cognitive science – "embodied-embedded cognitive science", which tries to put the otherwise disembodied brain back in the (biological) body and situate that body in the world in which it is embedded. Thompson also identifies these three kinds, referring to this last one as "embodied dynamicism" (Thompson 2007: 4). It is less obvious how such a theory could be of a kind with the classical and connectionist theories, and whilst Wheeler and Thompson do suggest this third kind is the beginning of something different, I want to suggest in this thesis that it is still the same as long as one condition remains – what Varela et al. refer to as "an already-given condition" (1991: 1):

> "A phenomenologically inclined cognitive scientist reflecting on the origins of cognition might reason thus: Minds awaken in a world. We did not design our world. We simply found ourselves with it: we awoke both to ourselves and to the world we inhabit. We come to reflect on that world as we grow and live. We reflect on a world that is not made, but found, and yet it is also our structure that enables us to reflect upon this world. Thus in reflection we find ourselves in a circle: we are in a world that seems to be there before reflection begins, but that world is not separate from us." (Varela et al. 1991: 1)

Cognitivism is not a "phenomenologically inclined" paradigm. (Post-cognitivism, by contrast, is). And yet, cognition, the very thing for which cognitivism enterprises to offer an account, is taken as an "already-given condition" in much the way that Varela et al. describe in the above passage. More specifically, it is taken as already-given in the cognitivist *model*.

There is a subtle distinction here about what exactly is being taken as already-given. In the above quote Varela et al. note that, phenomenologically, "minds awaken in a world", which is to say, something *is* already-given to the cogniser as they find themselves in a world. The important detail is that *what this* is, is not well-defined. All models of cognition take for granted and as already-given the basic condition of existence and that we find ourselves existing in a world. However, it takes cognitive work to figure out just what is going on from there – what matters and "what the problem" is, so to speak – but in cognitivist models of cognition, these variables are taken as already well-defined. In the post-cognitivist models, these are taken as not well-defined, and the "sensemaking" (§5.2.3) is something an entity has to do in order to arrive at a situation in which there is a clearly defined problem.

What, therefore, the cognitivist model, is unique in taking as already-given then, is not the *existence* of the cognitive situation, but the values of the basic variables – agent, world, and problem – which constitute the cognitive situation. This is something all three of the cognitivist theories do. When this part of the process is excluded from the picture, what is left to model is just how that situation is resolved, which is to say, how the problem is solved.

The theories thus account for the process of cognition once the existence of the agent, its world of information, and problem are "already-given". This is not necessarily a failure of the cognitivist model,

at least not on its own terms, but it is important for understanding how it develops a problem-solving conception of cognition.

3.1.2. Cognitivism in this Thesis: "Problem-Solving"

In the literature above on cognitivism in cognitive science, there are different headings that I am putting under this same banner. Classical, connectionist, and some variant of dynamical embodied, (and embedded) research typically distinguish themselves from one another, but in this thesis I include them all under the heading of "cognitivist" because they all share that important assumption which the cognitivist position makes, namely the "already-givenness" in their model of the cognitive situation. When this is taken as already-determined, there is not much left to explain except how that problem is solved by the agent, given the information it has. Again therefore, it stands to reason that cognitivism should come to a conception of cognition and intelligence as problem-solving.

Taking this view, then, when I speak of cognitivism here, I am referring to a paradigm of mind in the broadest possible sense, which is to say, *a way of thinking about mind and related concepts* like cognition, intelligence, and so on at a high level of abstraction and granularity. It is a "paradigm" of mind because whilst it includes different kinds of theories about these concepts, they are unified in their fundamental orientation on the question of mind – they are all looking at it in a similar way. My claim is that *the fundamental orientation of cognitivism conceives of "minds" in terms of "**problem-solving**"*.

The characterisation I develop in this chapter is neither arbitrary nor trivial. Later in this chapter I will argue that it affords a means of making sense of cognitivism's fundamental theoretical problems, the frame problem, the binding problem, and the grounding problem, not to mention the paradox at the heart of the xrisk narrative that superintelligent AI could be both superintelligent and super stupid, otherwise known as the orthogonality thesis.

To further substantiate this "problem-solving" conception of cognitivism and show that it is neither arbitrary nor inconsequential to conceive of it thus, I will speak about three paradigmatic concepts of cognitivism: "computation", the cognitivist model of cognition - what minds do; information, what is being computed; and "functionalism" – a framework guiding cognitivist inquiry. Each expresses and propagates a "problem-solving" orientation about mind, cognition, intelligence, and even consciousness, (though, it's unnecessary for purposes here to go as far as consciousness). An understanding of these three concepts and the thinking they afford is sufficient to grasp how the ideas involved in the xrisk narrative (superintelligence, orthogonality, and intelligence as instrumental rationality) are distinctly cognitivist in this way.

Again, I do not claim that this characterisation of the concepts is definitive or something to which all "cognitivists" would necessarily subscribe.

The rest of the chapter will discuss computation, information, and functionalism with a view to supporting the "problem-solving" characterisation of cognitivism. With these three fundamentals in place, a summary to tie it all together will be made by discussing the notion of a Turing machine. This will set up the discussion in the next chapter in which I argue that the claims of AI xrisk fall out of the cognitivist model.

3.2. Computation, Information, Functionalism

It is maybe not immediately obvious how "computation", "information", and "functionalism" frame, and are framed, in a way that generates a "problem-solving" understanding of mind, cognition, and intelligence, let alone what alternatives might be, or why it matters. In this section I will offer a brief description of how computation, information and functionalism are understood in the context of philosophy of mind and AI, and make the link to problem-solving from there. The key is that each of these concepts, in their own way, either assumes or require taking the cognitive situation as already-given.

3.2.1. Computation: Abstract and Physical

For the cognitivist, cognition is computation. The purpose of this section is to present an array of definitions of computation in the literature so as to get a sense for what the paradigm means by the term. Following Müller and Hoffmann (2017) and Piccinini and Maley (2021) I make a distinction between abstract computation and computation in physical systems. I begin by discussing abstract computation, placing it firmly in the heritage of Turing's (1937) work on computable numbers, and follow up with a discussion of Turing machines to give a working example of the classical conception of computation. Following the discussion of Turing machines, the questions that arise in the context of computation in physical systems will be discussed to get a more detailed and "concrete" picture of computation. With these conceptions in place, I offer a summary conception of computation, with a view to making clear how computation, for its part, both presumes and propagates a problem-solving conception of mind and cognition.

Piccinini and Maley (2021) offer a helpful overview of the dialectics and evolution of various positions on computation. They distinguish between computation in an abstract sense, as it is used by mathematicians, and computation in physical systems. I treat both in turn.

3.2.1.1. Abstract Computation: Computable Numbers and Turing machines

Computation in the *abstract* is an understanding of a process of transformation seeded and cultivated by Alan Turing in his famous (1937) paper. It is worth taking time and detail here. The paper was Turing's answer to the "Decision Problem", a problem proposed, along with others, by the mathematician David Hilbert at the 1928 International Congress of Mathematicians (Gleick 2012: 207). At the time, following the publication of Whitehead and Russell's three-volume *Principia Mathematica* between 1910-1913, and entangled with the cultural influence of the logical positivism of the Vienna Circle (Uebel 2021) and the linguistic turn in philosophy (Glock and Kalhat 2016), there was much concern with the foundations of mathematical logic (and, by implication, analytic philosophy). Hilbert's "*Entscheidungsproblem*" ("Decision Problem") asked whether there exists an algorithm for deciding, in first-order logic, whether a proof is derivable in that logic. Kurt Gödel had published his famous incompleteness theorem in 1931, which had important consequences vis-à-vis the Decision Problem, but:

> "Even though a particular closed system of formal logic must contain statements that could neither be proved nor disproved from within the system, it might conceivably be decided, as it were, by an *outside* referee – by *external* logic or rules." (Gleick 2012: 207) (emphasis mine)

This is to say, Gödel's work had not put all the nails in the coffin of Hilbert's hopes for establishing at least the *possibility* of firm foundations for mathematics. In his philosophical history on the origins of cognitive science, *The Mechanisation of the Mind*, Jean-Pierre Dupuy points out that "the incompleteness theorem did not quite provide a solution to the problem posed by Hilbert, however. For the theorem's import to be fully appreciated, it was necessary to clarify what was implied by such notions as "effective computation" and "finite procedure" ...in other words, what was still lacking was a rigorous, mathematical definition of the *algorithm*." (2000: 34; italics in original). Turing's machine (1937) presented tangible, mechanical definitions for these notions and so enabled a definitive response to Hilbert, showing that there exist undecidable propositions and that, therefore, there is not a universal procedure for deciding whether a proof is valid. But here, Turing's *question* is perhaps more interesting than his answer, and surely of more historical consequence. In effect, he asked Hilbert's question in a different way, a way that has since changed how "computation" is understood - he asked, "are all numbers computable?" (ibid, Gleick 2012: 207)

As is now becoming more remembered, computers were originally humans, and usually women. So, Turing's question had a very particular meaning. He was asking whether all numbers were computable *by humans*. As Gleick notes, Turing's "was an unexpected question to begin with, because hardly anyone had considered the idea of an *in*computable number. Most numbers that people work with, or think about, are computable by definition… Nonetheless Turing made the seemingly mild statement that numbers might exist that are somehow nameable, definable, and *not* computable" (Gleick 2012: 207) (emphasis in original). The "mild" way Turing defined computable numbers was as "those whose decimals are calculable by finite means", where "by finite means" acknowledged that human computers have only finite capacities of memory, recording, and time at their disposal with which to compute (Turing 1937: 230-231).

To show that a number was incomputable, and thus that there exists a class of proposition or proofs that are undecidable, Turing moved to show that some numbers are not calculable by finite means. This would be difficult to show as it pertains to humans because of that enigmatic gap between our inferences, gaps we bridge via the mystery of insight, intuition, and imagination and so on. So instead, Turing leaped an enigmatic gap of his own and considered the case of a "machine", stipulating that a number is computable if it can be written down by a machine. And in so doing, it is sometimes suggested that a Turing machine is even a new mathematical object, alongside numbers, sets, shapes, graphs, topographical spaces and the like.

> "We have said that the computable numbers are those whose decimals are calculable by finite means. This requires rather more explicit definition…For the present I shall only say that the justification lies in the fact that the human memory is necessarily limited. We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions…" (Turing 1937: 231)

Turing then proceeds to define this machine, describing what is now known as a Turing machine. A Turing machine has become the original and simplest abstraction of a computer and the notion of computation, the processing unit that computes information. He describes it as:

> …a machine which is capable of only a finite number of conditions…which will be called "m-configurations". The machine is supplied with a "tape" (the analogue of paper) running through it, and divided into sections (called "squares") each capable of bearing a "symbol". (ibid: 231)

The "machine" here is the processing-unit and the "tape" is the input. Strictly speaking the machine is not processing the tape but the "symbol" printed on a given section of it. The information in this case is borne by the symbol, not the tape. The tape is a delivery mechanism, or "substrate", that is all. (This will matter later on,[14] and particularly in the next chapter on post-cognitivist conceptions of cognition.)

Here computation is the process of the Turing machine processing the symbols on the tape. Turing describes what this looks like in terms of the behaviour of the machine:

> The possible behaviour of the machine at any moment is determined by the m-configuration…and the scanned symbol … This pair…will be called the "configuration": thus the configuration determines the possible behaviour of the machine. In some of the configurations in which the scanned square is blank (i.e. bears no symbol) the machine writes down a new symbol on the scanned square: in other configurations it erases the scanned symbol. The machine may also change the square which is being scanned, but only by shifting it one place to right or left. In addition to any of these operations the m-configuration may be changed. (Turing 1937: 231)

The computational process of a basic Turing machine is therefore sequential. First there is the "condition" or state of the machine ("m-configuration") and in that state the machine then reads off the symbol (or blank space) on the given square of tape (ibid). Together this is the "configuration" of the machine, state plus symbol. There may be arbitrarily many states, or as few as only three or four. In the first (Turing 1937) quote above, Turing states that in this model he imagines the machine to be "capable of only a finite number of conditions". In this configuration of state and symbol, the machine can then either "erase" or "print" a symbol, as well as "move" the tape. Like the symbols on the tape, the possible actions of the machine are discrete.

In sum, the basic computational process therefore involves the machine (processing unit/computer) moving the tape back and forth to read different symbols (input-processing) and either erasing a symbol, printing a symbol, or doing neither (output).

Now, an important detail at this point is that all this is described in a table which specifies the behaviour of the machine given its configuration (state + input) (Turing 1937: 233). A description of the behaviour of a Turing machine might also be called its "program". However, when we think of a program, the intuition is that the program determines the behaviour of the machine, whilst the description in Turing's

[14] See the discussion about "substrate-independence" in section §3.2.3.4. on Marr and the Tri-Level Hypothesis.

tables are simply descriptions, and do not influence the behaviour of the machine. The line of this distinction is important, amounting to the question of whether the term "program" is something ontological, or whether it is just an epistemologically useful way of describing the behaviour of the machine. In the next section on computation in physical systems it is important to distinguish between physical systems, like a rock or planet, which can be *described* as performing computational transformations of states, but of which we would not want to say that they are actually computing in a cognitive sense.

This then is a rudimentary illustration of a Turing machine. Certain things have been omitted. For example, the input need not be digital/discrete but can be analog/continuous. I also left out the matter of the "memory" of the machine. These and other details are important when it comes to a evaluating a Turing machine as a comprehensive model of mind, but a sufficient grasp of *computation* in the abstract can be had without them. Again, I am using Turing Machines to illustrate in the simplest way possible the foundations of the computational process so as to offer a sense for the problem-solving picture of cognitivism.

Having said that, one "detail" that *is* an important complement to a basic understanding of Turing machines is the notion of "Universal Turing Machines". Turing addresses this later on in the same paper (1937: 241). It is an important part of how Turing machines show that there is not an algorithm which can decide, for all theorems and proofs, that they are derivable in their own logic.

A Universal Turing Machine is a Turing machine which is capable of simulating any other Turing machine. The way Turing puts it is that "it is possible to invent a single machine which can be used to compute any computable sequence." (1937: 241) Because the behaviour of any Turing machine can be fully described by a string of symbols, it is possible to use that string as input for the Universal Turing machine, which can in this way be "programmed" to compute according to the very rules of the first Turing machine.

This recursion is vital to the Decision Problem. Turing's ingenuity was to show how a program can refer to itself such that a (Universal) Turing machine can take its own program as its own input, leading to contradictions. Gödel employed the same moves in his work on incompleteness. Indeed, by all accounts Turing learned this from Gödel's work. In the context of "deciding" whether a statement is decidable in its own logic, when a program for deciding runs, this self-reference generates an infinite loop, meaning the program never "halts", and a decision cannot be reached (Mitchell 2009: 56-68). The details are basic but humbling for their capacity to bewilder and confuse. There are famous, non-technical examples of the same thing – paradox – in philosophy. Amongst the oldest in the Western tradition, going back to the New Testament, is the Epimenides paradox (Sorensen 2003: 93). It is enacted when the statement "All Cretans are liars", *is uttered by a Cretan*. (Epimenides was a Cretan). The statement is only paradoxical if stated by a Cretan. A more contemporary example is "this sentence is false". In both cases, trying to determine the truth of the statement sends us into an infinite loop of contradictions - if it is true, then it is false, but then it is true, which means it is false…and so on. These older paradoxes express the detail highlighted by Gödel and Turing that some statements are not decidable *in their own logic*. Something external is required to prove the statement.

There is one last thing to note which is not normally discussed in these contexts, but which foreshadows some of the distinction between cognitivism and post-cognitivism to come. Regarding the Epimenides paradox, I said it is *enacted* when stated by a Cretan. Here lies a note of distinction between the cognitivist paradigm and the enactivism of the post-cognitivism discussed in the next chapter. The statements are *made* paradoxical, the paradox is *enacted* because it is what the statements *do* that is inconsistent with what they are *about.* The Epimenides paradox is enacted by what the Cretan is doing.

What he *does* in saying the statement, is inconsistent with what the statement is about, its content. "Is-ness" and "about-ness" are shown to be inconsistent in these paradoxes. René Magritte's famous painting *The Treachery of Images* enacts the same paradox, recognising that it is a picture of a pipe, "about" a pipe, but the picture *is* not a pipe. The distinction between is-ness and about-ness is a simple way of recognising that a computer computes models that are about the world, rather than "is" the world. A model of a hurricane isn't "wet" and so on. As it concerns this thesis and whether AI could become more ethical than humans, the possible challenge is the following question – if AI can solve a model or representation of or "about" a moral problem, can it perform or enact a solution to the actual problem itself?

In any case, the universal Turing machine is significant for a simpler reason too. A regular Turing machine might seem quite limited in its capacities, being able to perform only a narrow range of computations the way a calculator can do arithmetic and so on, but nothing else. In this way, a sceptic might be tempted to say that whilst a Turing machine is a helpful model of the fundamentals of computational cognition, there are some things it won't be able to compute that, say, biological organisms can. One thing the notion of a universal Turing machine seems to show is that such a machine can in principle (be made to) perform any computation that any other Turing machine can do.

Moving from the processing unit to the processing itself, there are several important things to note again about the process of computation in this abstract model. One is that it is sequential and linear, in two spatial dimensions. Depending on the "configuration" (state + symbol) the machine can only move the tape in two directions. The sequential character to computation means that the computer/processing unit is processing "one thing at a time". This is turn speaks to the second thing – the tape is divided into discrete "sections" or "squares" which bear the symbols, and the resulting states and movements of the machine are then also discrete. What is significant and relevant for present purposes is the way in which the process of computation involves *well-defined states and symbols and input.* This leads to perhaps the most interesting and relevant thing about a fundamental look at computation.

It is what is implicit and assumed in the image I have presented. In the cognitivist model, the information which comes in, is as an *already-defined* variable. The symbols are taken to pre-exist, or be pre-defined, *before* they are processed[15]. To say that the information is "already-defined" is to say that the variable is already defined. The meaning of the symbol in the cognitivist model, the particular significance it endogenously takes on for and by the agent, is taken as already-given. The processing unit does not, nor need not, define the information which comes to it for processing, or the meaning thereof. That is, the processing unit does not participate in defining what the information means, as we will see is the case in the post-cognitivist models in chapter five in which agents endogenously "bring forth" the meaning of what is encountered.

The difference is that in these post-cognitivist models, the agency of the agent is involved in generating *the basis* upon which it interprets meaning. By contrast, in cognitivist models, the information and symbols a Turing machine encounter can take on particular meaning for it, but a given Turing machine is not involved in generating the basis upon which it interprets what it encounters.

For the sake of intuition, consider for example the way in which a child progresses through (Western) institutional structures of education. The child begins by adopting the interpretations offered to them

[15] This applies to analog information as well as digital. The digital-analog distinction concerns the metaphysical character of the information, smooth and continuous, or definite and sharply bounded, but this does not necessarily effect the way in which it is cognitively processed by an agent. A different "processor" may be required, but the cognitivist model in which the input in its already-defined form, still applies. That the meaning of the information is pre-defined is unaffected by whether it is analog or digital.

by their teachers, but, we hope, eventually comes to develop not only their own interpretations of *Othello* or *Catcher in the Rye* or something else, (a step which still takes as already-given that such works have importance, as defined exogenously by teachers), but they also come to develop for themselves their own sense of why such works are, or are not, important such that their interpretations are grounded in their endogenously generated basis of meaning. It is this kind sense of *endogenously* discerning meaning that is key to the distinction here between the cognitivist and post-cognitivist models.

How to understand what is being said here takes us in fact to the place of the very problem-solving – problem-defining distinction on which this thesis is based. A Turing machine reads a symbol on a length of tape, but, as humans, we often do not know what we are looking at, and, to that extent, we do not know how to respond to what we are looking at. The particular significance or meaning remains to be determined. Sometimes we have to work to discern what something is or means in terms of a behavioural response, both when it is abstract, for e.g. trying to discern what my words here might mean, or physical, trying to discern what the incoming object in the jungle is. A Turing machine does not have to work to figure out what the symbol is and how to respond, whilst often human perception very much involves having to do this. This is an epistemological point. The incoming tiger in the jungle is presumably an ontologically well-defined phenomenon even if we can't make it out, but the information that presents itself to us can be ambiguous and ambivalent. In contrast to the Turing model of computation, what things seem to be, and how to respond to them, are often not clear for humans.

For the sake of further clarity, consider first the phenomenon that two people can look at the same thing and it can mean very different things to each of them. It can take on different significance for them, be it a person – a loved one for example – or a photo of a loved one. Recognising this difference is all that is required to notice that whilst the basic stimuli or information[16] coming in may be identical, their effect is not. The meaning or significance of the stimuli is not pre-established for any given agent the way it is for a Turing machine. In the case of loved ones, the significance might be more or less immediately presented in awareness, but sometimes that significance is not so immediately forthcoming. If a red irritation appears on the skin of my leg, the significance of it is not immediately clear. I don't know how exactly to proceed. I don't know what it means for me really until I can be sure of whether it is an allergy, an insect bite, an abrasion, or something else.

One last way to put the difference is to say that the model of computation, the fundamentals of which are expressed in the model of a Turing machine, assumes capacity to read. A Turing machine is already capable of reading that which is presented to it. Human children, though, are not, both in the literal sense of reading, but also in the metaphorical sense of reading and endogenously discerning the significance of what they encounter. In the case of literal reading, it takes time before a child even recognises that particular arrangements of lines on a page or screen are letters and words and so on – that they are things worthy of attention beyond mere edges in their visual field – and even then it takes more cognitive work and development to recognise, to discern, and encounter any meaning in the words. The same can be said about regular, non-symbolic objects in the world, from the tiger in the jungle to a bus on a street. Even our phenomenological experiences require a reading capacity of a certain sort, and this "reading", literal and metaphorical is not already-given to humans as cognitive agents. A Turing machine though, can always already read what it encounters. It is in this sense that what a processing-unit computes is, in this model, already-defined and already-given for it.

[16] Discussions of information do not usually make this distinction between a basic stimulus and the meaning or significance of that stimulus for an agent.

This may turn out to be legitimate way of modelling cognition, if it is the case that things "out there" are already well-defined or real before an agent comes to interact with them, but it is an assumption the cognitivist model makes.

The implications for questions of realism are a place of significant divergence and nuance between cognitivist and post-cognitivist models of cognition but are not, alas, the focus of this. It bears keeping in mind thought that work on AI, being rooted in the cognitivist model, rest on considerable assumptions of realism. See Varela et al. (1991: 164-172) for an insightful discussion.

Whilst important, the realism question is not in fact the most significant point for my purposes in this thesis. Instead, it is the way in which the already-givens of computation begin to lead to a problem-solving view of cognition.

Consider some rhetorical questions, beginning with "why this particular information, why these particular symbols?" The fact that these variables are already-given reveals to us that computation, as a model, does not involve on the part of the agent an arbitration, discernment, "reading" or definition of the information and its meaning. The computer simply has to deal with whatever it is fed, whether symbols or something else. Whilst the computer *does* respond to its input in a way that causally and counterfactually makes a difference to what it encounters later, the procedures for how to read and respond are *also* already-predefined, something that is an open question in the enactivist model I discuss later in the thesis.

Thus, whilst now it would be too soon to say just yet from this that computation therefore consists solely in "solving" things, such a point at least begins to come into view.

Again, this is based on two specific "already-givens" in the computational model of cognition just discussed. Firstly, the meaning of the information (for the Turing machine) is something already-given in the model and so is how to respond to it. Further, the existence of the Turing machine itself is also already-given in the model. It might be silly to call out a model of cognition for taking as already-given the thing which is cognising but, in the post-cognitivist, Enactivist model, accounting for the genesis of the cogniser is a vital part of the model of the cognitive process and is a major part of accounting, in a non-arbitrary way, for how the cogniser encounters the meaning that they do. Together this means that, in the cognitivist model, the machine along with the information and the meaning thereof, (for the machine), are all taken as already-given in a way that begins to illustrate how the restrictions of the model lead to a description of a problem-solving process.

There are significant philosophical consequences to these assumptions. If the model takes the meaning (to the agent) of the information as already-given, then that meaning is treated as arbitrary. What we end up with then is, rather than a model in which we try to understand *why* something is meaningful to an agent from their perspective, a central feature of human personal and social life, is a model in which we take the meaning for granted and model only *how* an agent executes a function of already-given, and arbitrary significance.

This is a central distinction in the discussion to come on the notion of information. It also shows up in the following chapter in which I lay out how these assumptions yield the ideas of AI xrisk idea like the Orthogonality Thesis. In the end it *is* a powerful way to model cognition because it is the lowest common denominator across entities and so enables us to model machines and non-human animals alongside humans. Then again, it is also the *lowest common denominator*. The feature is also a bug in the sense that it does not necessarily include everything that is worth including in a model of human or natural cognition.

This concludes the presentation of computation in its abstract sense. The fundamental process expressed in the model of a Turing machines offers a foundational understanding of computation, but when it comes to talking about some *physical* systems, it does not always seem to fit. In the next section I turn to this matter.

3.2.1.2. Computation in Physical Systems

A model of computation in the abstract describes a particular process of transformation, and whilst systems natural and artificial do perform and undergo transformations, philosophers recognise that not all these should count as computation. The question of computation in physical systems is largely a matter of drawing this distinction. Chalmers (2011) describes it as the question of, "When does a physical system compute?", or sometimes as the "problem of implementation" (see also Chalmers 1994):

> "The mathematical theory of computation in the abstract is well-understood, but cognitive science and artificial intelligence ultimately deal with physical systems. A bridge between these systems and the abstract theory of computation is required. Specifically, we need a theory of implementation: the relation that holds between an abstract computational object (a "computation" for short) and a physical system, such that we can say that in some sense the system "realizes" the computation, and that the computation "describes" the system. We cannot justify the foundational role of computation without first answering the question: What are the conditions under which a physical system implements a given computation?" (Chalmers 2011: 327)

The question at the end is arguably the question which has motivated the various theories of computation in physical systems. As (Piccinini and Maley 2021) point out, there are systems which can be *described* in computational, input-output language, but which we wouldn't want to say are cognitive systems. "Does a planet compute its orbit?" A planet in orbit undergoes state-transformations that are described by the principles of relativity, all in much the same way that a computational system undergoes state-transformations that are described by an algorithm or program. And yet, a theory of cognition as computation which failed to distinguish computation in a planet from computation in a human would seem to have missed something distinctly cognitive.

> "Our question may be reformulated in the following way: *which* input-output processes, among the many exhibited by physical systems, deserve to be called computational in the relevant sense?" (Piccinini 2016: 209)

What the case of the planet points to is that the physical transformations realised by planets, and other such objects which we would not consider as minds, are not identical with computational transformations. The challenge therefore is to articulate an account of when this relationship *does* exist, when computational transformations and physical transformations have a non-arbitrary, non-coincidental, and non-random relation. This line marks the boundary between systems we take to be

obviously cognitive, like humans, and systems that are not obviously cognitive, like planets. There are some systems that are on this boundary though[17].

In any case, Piccinini and Maley (2021) summarise four different kinds of theory of computation which variously circumscribe restrictions on what kinds of state-transformations can count as a genuinely computational transformation: "mapping" accounts; "causal", "counterfactual", and "dispositional" accounts; semantic accounts; and syntactic accounts. Piccinini and Maley (2021: 5) use a diagram to explain the problem:

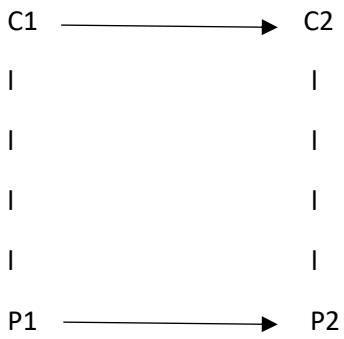


Figure 3. Diagram of Relations Between Computational States and Physical States

Here they identify some "mapping" relationship between transformations of physical states and computational states and take in need of explanation the dotted lines between C1 and P1, and C2 and P2.

The first account, the "mapping" account, says that transformations in computational states, as specified in a "machine table", "map" onto the physical transformations (Putnam 1960, 1967, 1975). One problem for this account which quickly presented itself was the "productivity of thought" mentioned earlier in the discussion about Representation. The possible physical transformations seem to be finite, whilst the possible transformations of thought, and thus computation, seem to be infinite, so saying that computational states mapped onto physical states was not precise enough. Further, as (Piccinini and Maley 2021) point out, this still allows that every physical system could be implementing computation, as suggested by (Putnam 1988 and Searle 1992)[18].

> "…it is perfectly possible to build computational models of many physical processes, including respiration, digestion, or galaxy formation. According to the mapping account, this is sufficient to turn lungs, stomachs, and galaxies into computing systems, in the same sense in which calculators are computing systems and brains may or may not be. As a consequence, many things – perhaps all things – are turned into computing systems…given the mapping account, not only is everything computational, but everything also performs as many computations as it has computational descriptions." (Piccinini 2016: 208)

[17] These are often distributed systems, like forests, or multi-agent collectives of animals. So far, computation is only used in a descriptive sense for these systems, but Krakauer et al. (2020) show in information-theoretic terms that these systems can be individuated in a way that is not a million miles from "obviously cognitive", even if not entirely there on the spectrum.

[18] There is a notable connection to the Frame problem here, (see §3.4.1.) of this chapter for a more in-depth discussion), namely, how do we specify – i.e. "Frame" – which mappings are the relevant ones if we clearly have an idea that amongst many, only some subsets are relevant ones?

The Causal, Counterfactual, and Dispositional accounts are attempts to place restrictions on acceptable mappings so as to avoid this problem. They still speak in terms of mappings, with an additional concern about the metaphysical nature of lines in the diagram in Figure 3 above (Piccinini and Maley 2021). These accounts are cases of "it does exactly what it says on the tin". The causal says that physical states implement computation when the transformation between P1 and P2 is causal, and that the computational transformations map onto that causal transformation (Chrisley 1994, Chalmers 1994, 1996, 2011). The counterfactual only permits mappings to counterfactually stable transformations between states. As for the dispositional account, the mapping of computation to physical states is genuine computation if, in transforming from P1 to P2, the physical thing manifests a disposition (Klein 2008).

One thing is worth noting before presenting semantic and syntactic accounts of computation. The three above accounts are all attempts to make sure that there is some non-arbitrary mapping between possible transformations.

Turning now to the semantic account of computation. This position was introduced by Jerry Fodor (1975, 1981), and is the basis of Fodor's famous quip that there is "no computation without representation" (Fodor 1981: 180). The idea is that computation is the processing of representations, and representations are the semantic content, what the symbol processed by a Turing machine stands for. So, whilst a rock or planet may be described in terms of input-output transformations, the semantic account says that it is not computation because they are not computing representations. The challenges for the semantic account include how to individuate representations such that they can be meaningfully distinguished so that we know what representation is actually being processed. There are internalist and externalist positions on this. Another important question concerns what counts as a relevant kind of representation. Fodor, in his famous *Language of Thought* (1975) argued that semantic representations have to have a syntactical structure. Recall the "John", "James", and "loves" example from the discussion of Representation earlier. It is clear that each of these words or symbols semantically represents something, but taken in isolation the symbols don't mean much. A syntactic, structural relation – subject-verb-object, say – between the words brings out the particular semantics of each symbol. The syntactic account is therefore a genre of semantic account[19].

Lastly, there is also mechanistic account of computation (Piccinini 2007, 2015). The reasons in favour are nuanced but, interestingly enough speak to the "already-given" critique of cognitivism that has been raised in throughout this chapter. The mechanistic account takes a functionalist stance on the individuation on "vehicles" of "concrete computations". This is to say, it takes a functionalist stance with regard to individuating computations. An immediate advantage to this approach, and notable particularly in the context of this thesis, is the way a function "frames" the computations, giving them particular meaning without having to resort to semantic or syntactic accounts. Recall from the discussion of Turing machines earlier the question of "why this particular computation?" In that abstract account of computation, why the tape had the symbols that it did, and why they were processed according to the rule that they were, were all "already-givens". Piccinini's mechanistic account addresses this by saying that the particular symbols, and rule, are those relevant for the system of mechanisms to process, given their function. The functional approach brings a teleological frame to the computation which has the effect of specifying the relevant computations (Piccinini 2015: 10). Lungs

[19] One last question for the semantic account is what gives representations their semantic content. This is a slightly more metaphysical question. Some, like Dennett in *The Intentional Stance* (1987), and Egan (2010) take an Instrumentalist view, in which ascribing semantic content to representations is "heuristically useful" (Piccinini and Maley 2021). Then there are Realists, some of whom are also Naturalists, and some of whom are not (see for e.g. Fodor 2008 for a discussion of the relationship).

pump air, brains process electrochemical potentials, computers process code, humans process all sights, sounds, touch, etc., and in each case the kind of computation may be different – analog/digital, distributed/serial and so on. In this regard, another advantage of the account is what he calls "medium-independence".[20]

As for the "which account" question, a pluralism seems to be popular. See (Dale 2008, Edelman 2008). For the purposes of this chapter, it is the character of their fundamental similarities I wish to foreground, so it is unnecessary to arbitrate between them.

Having now presented how computation is understood in abstract and physical systems in cognitivism, an attempt will made to summarise a broadly cognitivist conception of cognition in a definition. Again, more than what is said, what is left out is most informative, with respect to distinguishing cognitivism and post-cognitivism. What is left out is what leads to the problem-solving conception of cognition.

3.2.1.3. A Cognitivist Conception of Computation

A few quick comments on the title of this subsection: the title says "a" because it is one and not necessarily "the", if such a thing is possible; "cognitivist" because it aspires to speak to the variety of cognitivist theories and distinctions just described in the previous two sections on abstract and concrete computation. The definition is not meant to synthesise all the definitions though it is meant to be inclusive. It is meant to offer a coarse-grain and working impression of cognitivist conceptions of cognition. First, I will describe the things a cognitivist account of computation seems to need to include, then the things it excludes will be noted.

We can begin with the following definition:

Summary definition: Computation is *rule-defined information-processing.*

This is general enough to encompass the abstract notion of computation described in the section on Turing machines, as well as offer a means of distinguishing computation in physical systems. Again, it is not intended to be a definitive account. It is intended to be representative of cognitivism.

It includes abstract computation in the following way. There is some processing of tape carrying symbols, and the particular behaviour of the processor is expressed according to some definable rule. The rule can be a description, as in Turing's "machine tables" with no actual influence on the behaviour, or a program, in the conventional sense, which dictates how to process input. For the purposes of the thesis I am keeping this as an open question because, particularly for physical systems, and even more so for living systems, it is a moot point for many physical systems whether the rules by which we can describe them are programs or descriptions. As we saw above though, there do exist various accounts of concrete computation which do specify various kinds of rules and conditions according to which something counts as computation: mapping, causal, counterfactual, dispositional, semantic, syntactic, and mechanistic.

[20] See again §3.2.3.4. on Marr and the Tri-Level Hypothesis for more discussion of medium-independence.

"Information" is also suitably ambiguous for a definition of computation at this altitude. It suits the abstract, Turing machine conception of some sort of symbol processing. With a suitably broad interpretation of information, it also works for concrete computation, allowing for the varieties of input particular to different systems.

"Rule-defined information-processing" as a summary definition of computation thus includes at a high level of abstraction at least most of the variety of things taken under the heading of computation.

With the computation piece in place, we can now turn to the notion of "information" – that nebulous thing computed. Information and computation are intimately entangled (Piccinini and Scarantino 2010, 2011). It is hard to have one without the other. In the next section, I go on to show how information also leads to a problem-solving conception of cognition. I will finish this section with a quote from Dupuy's (2000) history of the origins of cognitive science: "This, at least, is my reading of what motivated the pioneers of cognitive science: the notion that thought, mental activity, this faculty of mind that has knowledge as its object, is in the last analysis nothing other than a rule-governed mechanical process, a "blind" - might one go so far as to say "stupid"? - automatism." (Dupuy 2000: 39).

3.2.2. Information: Control and Communication

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

- T.S. Eliot

In the context of the computational theory of mind, and the cognitivist paradigm of mind of which I am claiming it is the most distilled expression, information comes with a lot of metaphysical baggage. It did not always have this baggage. At least, when Claude Shannon was quantifying and formalising it into Information Theory in the mid-20th Century, he was clear that he was not intending his formalisms to be interpreted in any but a concrete way appropriate for the purposes of an engineer. His concerns were those of an engineer, not a philosopher. The meaning of information has travelled far since Shannon, further even, perhaps, than from the much older discussion of the in-forming of marks on a wax tablet in Plato's *Theaetetus* (2015: 190e5–196c5).

In this section, my aim to is to present information as a second and fundamental concept of the cognitivist model of cognition. I try to show that as a working feature of the model, information relies on the same fundamental assumptions as computation and, as I show in the following section, functionalism. Again, the most basic assumption is that information, as it is used and understood in the cognitivist model, only begins to work once the cognitive situation is taken in the model as already-given. Information is typically how a model of cognition that describes the phenomenon in terms of computation is able to speak of the computations *meaning* something.

From the perspective of this thesis, it is at this point that the mess begins. There are at least two ways, in the context of models of mind, in which talking about information *meaning* something can be confusing. The first is well-recognised and consists in the fact that there are at least two main senses of what "meaning" here amounts to. One sense is quantitative and is the sense that is dominant in cognitivist conceptions of mind. It is this sense on which I focus what follows. This sense of meaning of information, if it can be referred to like this, is defined rigorously, which is to say mathematically. It

is the theoretical basis of most, if not all, computational hardware and software systems, including contemporary AI systems. As such, it is important to be aware of the way in which it is distinctive of the cognitivist model of mind. One of the main things I want to do in this thesis is bring awareness to how we do think about artificial intelligence and the palette of possibilities which are available. This quantitative sense of information is the theoretical basis of most of our thinking about AI, and it is deeply rooted in the cognitivist model.

Standing next to the quantitative conception of information and how it "means" something is the sense of meaning in the more everyday sense, one that despite this familiarity, or perhaps because of it, is comparatively difficult to pin down in a rigorous manner. It is distinctly more phenomenological, ranging from the everyday to the existential; qualitative where the other is quantitative. If the quantitative sense of information involves a question of "how *much*", the qualitative is more "why". Thus, these two ways of talking about the "meaning" of information is usually the first place where things get messy. The rest of this section is articulated around this distinction, one I refer to as the "split".

Another place, and less familiar in the discussion, concerns the question of "meaningful to whom?". Here the distinction concerns the difference between the modeller and the modelled. The crucial assumption I have cited in the cognitivist model is that it takes the cognitive situation as already-given. What this means here is that the 'meaning' of anything in the situation *for the modelled agent* is already-given, defined *by the modeller*. When meaning for the modelled agent, (in the qualitative sense of meaning,) is already there, then it is not something for which the model has to account. This is to say, the model is not then a model of a process involving qualitative meaning for the modelled agent. This aspect is implicitly included by the modeller when they interpret the model. Instead, the model is a description of how a modelled agent does something – solves a problem, achieves a goal – that is qualitatively interpreted by the modeller. I point to this distinction throughout the thesis when I emphasise meaning as the *endogenously* generated significance of something for an agent, from their perspective.

As I now turn to talk about information and the role it plays in the cognitivist model of mind, these two distinctions – that between the qualitative and quantitative senses of information and that between meaning being endogenous to the modelled agent and meaning being projected in by the modeller – are key to keep in mind to avoid confusion. I begin here now highlighting the basic divergence at the heart of the notion of information, and how cognitivist models work with one side of things.

In more detail now, the divergence consists in the fact that there are now at least two senses of "information": Claude Shannon's, which concerns the *quantity* of information in a message, and the more common semantic sense, which concerns the *meaning* of information, that is, the particular significance an agent endogenously encounters in their particular situation. Both senses are relevant and are at the heart of the discussion in this section.

A useful point of departure for making sense of the place of information in cognitivism's problem-solving orientation is Gregory Bateson's suggestion that information is "the difference that makes a difference" (Bateson 2015). The first record of the phrase is in his 1970 Alfred Korzybski Memorial Lecture, titled "Form, Substance, and Difference"[21]. Bateson's definition contributes to the developing picture of Cognitivism here because the focus on "difference" opens up the discussion in a way that can make sense of the *split* between these two senses of "information" at the centre of things.

[21] Korzybski is the mind behind the insight that "the map is not the territory".

With the publication and emergence of Shannon's work in 1948, "information" was explicitly disassociated from "meaning". Shannon's mathematical work concerned *quantities* of information. Around the same time, important interactions were occurring across disciplines of mind research otherwise removed from each other. This happened most famously at the Macy Conferences between 1946-1953[22], which were attended by Bateson and Shannon, amongst many others, and were the birthing ground of the field of Cybernetics (Wiener 2013; Gleick 2012: 242-248). As far as information is concerned, Cybernetics was important because it was the closest thing to a bridge across this "split". It aimed to offer, in the language of Shannon's information theory, a picture of how mechanical systems were related to autonomous systems of nested feedback loops, systems which, as they become more elaborate, begin to resemble living systems. For these systems, it begins to make sense to speak of information in the sense associated with meaning, and not just in terms of measurable quantity and so the split between the two begins to be brought together.

As well as this broader collective Cybernetic movement, the individual work of Turing also bridges this split in 1950 with his famous "Computing Machinery and Intelligence Paper". Through Bateson to Shannon, Wiener, and Turing, I will attempt as I go along to render a working picture of information as it is currently understood and operational in AI and philosophy thereof. The aim in the end is to show how information too supports a problem-solving characterisation of cognitivism.

Having suggested that there are at least two major conceptions of information, the important question, and perhaps the most immediate, is "well, which definition is actually used then?"

It seems that on the theoretical side of cognitivist mind research, attempts are made to keep in view both sides of the split, but on the engineering and computer science side, the side concerned with actually building AI systems, it is the legacy of Shannon's conception which is in operation. For the purposes of this chapter and thesis, and an attempt to characterise the cognitivist paradigm, both sides of the split are important. As well-defined and articulated as particular definitions of information can be, there seems consistently to be some degree of smudging going on to cross the "gap"[23] between, on the one hand, functional machines for which information, in its quantitative sense, manifests physically merely as means of effecting control and coordination of the machine body, and on the other side, for humans, for whom information seems to be something both obviously and ambiguously more.

3.2.2.1. The Difference that Makes a Difference

Bateson knew Shannon. Having attended the very same Macy Conferences, Bateson was well aware of the contemporary conception of information in his time (Shannon's) (Dupuy 2000, Gleick 2012). Some twenty years down the line from the Macy Conferences which were to yield the Cybernetic movement, Bateson (1970) presents his conception of "information":

> There are differences between the chalk and the rest of the universe, between the chalk and the sun or the moon. And within the piece of chalk, there is for every molecule an infinite number of differences between its location and the locations in which it *might* have been.

[22] The Macy Conferences ran from 1941-1960. The focus on what was to become Cybernetic systems seems to have occurred mostly at those conferences between 1946-1953 in which discussions focused first on the dynamics of homeostasis, then reflexivity, then on "virituality" and emergent behaviour (Hayles 1999: 16).

[23] There is a relation here to the "explanatory gap" in philosophy of mind between measurable matter and phenomenological experience, one worthy of further exploration for the ways in which they historically and genealogically are related.

> Of this infinitude, we select a very limited number, which become information. In fact, what we mean by information—the elementary unit of information—is a *difference which makes a difference…*
>
> But what is a difference? A difference is a very peculiar and obscure concept. It is certainly not a thing or an event. This piece of paper is different from the wood of this lectern. There are many differences between them—of colour, texture, shape, etc. But if we start to ask about the localization of those differences, we get into trouble. Obviously the difference between the paper and the wood is not in the paper; it is obviously not in the wood; it is obviously not in the space between them, and it is obviously not in the time between them. (Difference which occurs across time is what we call "change.") A difference, then, is an abstract matter. (Bateson 2015) (emphasis in original transcription)

In order to unpack Bateson's view here, there are some important things worth noting concerning the condition of possibility of differences that make a difference. There is no information in randomness (ubiquitous difference) or in homogeneity (pure identity/no difference). Differences cannot make a difference if there is only sameness and no difference (homogeneity), nor can differences make a difference if there is not enough sameness (randomness). There must be enough pattern (sameness) such that a difference can make a difference. There has to be a certain amount of stability and identity from which differentiation can occur. Information in this light is some pattern of difference, differentiating itself as such, against a background.

This shows up in the everyday, but now interesting, context of jokes of the "3 men walk into a pub" variety. Three is the minimum quantity required to establish *and then break* a pattern. Hence, the first man walks into the pub, does something ordinary, the second man repeats it, creating a pattern, then the third goes in and breaks the pattern to humorous effect. In more formal discussions of information, the notion of "surprise" is often invoked to name this difference. So, we can say that the difference that makes a difference is one that breaks a certain pattern. In the next chapter, in section 4.3.2.2. Complexity and Constraint Closure, this pattern-breaking is explored in terms of the notion of "symmetry-breaking", when there is an asymmetry or "non-equilibrium" dynamics across the threshold of the boundary of an organism. In information theoretic terms, such pattern-breaking and symmetry-breaking can be used to individuate agents or collective agents even (Krakauer et al. 2020). Pattern-breaking is thus humble, but of profound significance.

Going into a bit more detail now, a key thing to bring out is that the difference which makes a difference is one of many differences. Information is not just "difference" tout court but a difference *amongst* differences – information is *one that makes a difference*, a different difference. This prompts some questions: how is it that some differences make a difference for a given agent and others do not?

What we can recognise at least is that not all differences make the same difference to one agent, and not all differences make the same difference to different agents. Indeed, some differences make more of a difference for an agent than others. We might say they are more relevant differences in this regard. We could define differences to be relevant for an agent *to the extent that, and in the way in which*, they make a difference to that agent. Abstract and physical things make differences in different ways, and, often, things which are closer, literally or figuratively, tend to make more of a difference than things which are far away. What is relevant to one agent is not always relevant in the same way or to the same extent. This means that the same objective feature of an environment can take on very different

meaning. A picture of a deceased relative presents the same feature for two agents but makes different differences to the one for whom it is a picture of their grandmother. This is all to say, information is not just any difference, but a particular difference, one which effects particular meaning and significance to an agent.

Two important questions in this direction are: what, then, are the conditions in which a particular difference encountered by an agent makes a difference to that agent? And how do we explain why it was *that particular difference* that made a difference for the agent?

The matter of accounting for "why *that* difference" in the world makes a difference to a particular agent reveals something important about information as a fundamental concept of cognitivism, something that can be understood via the "Frame Problem" (Dennett 1984). The frame problem was mentioned in the introduction as one of the fundamental problems for cognitivist cognitive science and is explored in more detail in §3.4. Here, a working understanding goes a long way. The frame problem is the problem of how to specify, for an agent in a given context, how and why it is that what is relevant to them, from their perspective no less, is relevant in the way that it is. In this regard, the frame problem might generally be framed as *how to specify which difference makes a difference* for an agent, *and from their perspective.*

Above Bateson says that "within the piece of chalk, there is for every molecule an infinite number of differences…Of this infinitude, we select a very limited number, which become information." This act of selection amongst possibilities (which is central to Shannon's theory too) effects/is the effect of, a particular framing on the situation such that particular patterns are observed and that particular differences are informative.

To see this, consider the question "what is the function of a screwdriver?"[24] Or, put another way, "what difference does the screwdriver make for an agent?" Obviously, in principle, a screwdriver is "for" turning-in screws. However, it also makes for an effective paper-weight, a grim, violent stabbing tool, a convenient head-scratcher or door-stopper, an imaginative microphone, an improvised conductor's baton, an artist's paintbrush, and even an abstract object for making ironic sense of something as abstract as information. In each case the difference the screwdriver makes for the agents involved is different. That is, what seems to determine which difference makes a difference for an agent, is necessarily contextualised. What counts as information emerges in a particular perspective or frame on a situation. In the terms of Shannon's theory to be discussed, what counts as "signal" and what counts as "noise", can be quite different even if we are all talking about the "chalk". It depends on how the frame is *defined* for and by the agents involved.

Bateson's intuitive conception of information offers a lot of purchase on a notion of information. It is a suitably general sense of the term. Putting it together with computation, we can say that a computational mind processes the differences it perceives, the differences that make a difference to it with respect to the function or program it is executing. This understanding is a casual and not technical characterisation, but nonetheless appeals to the intuitively epistemic character of information, informing or making a difference *to* someone or something. In this way it is more obviously connected with theorising about the nature of mind and cognition than something as abstract as information.

[24] The credit for this particular example belongs to Stuart Kauffman (1993, 2019).

However, the more formal, operational, and famous account of information as far as computational systems are concerned, is Shannon information. It is not so obviously connected to a theory about mind, which stands to reason as it was developed in the context of engineering problems, but it has nonetheless become the go-to account of "information" in the context of AI. Whilst Bateson's intuition-led definition of information does not necessarily lend itself to a problem-solving account of cognition, Shannon's conception does, and it is the conception that is wielded in cognitivist AI work. Bateson's is nonetheless a valuable introduction to the notion of information because, apart from being an important character in the history of the idea, his view of information in terms of "difference" encompasses information in both the senses that is has come to be known. Namely, as a quantitative measurement regardless of its meaning, the sense we find in Shannon information, and the more qualitative sense of information that involves the "meaning" of the information regardless of its quantity. Following Bateson's take on information, the split between these two senses of information helps to make sense of how information has come to be understood mostly in Shannon's sense in philosophy of AI, and how Shannon's sense of information supports a problem-solving view of cognition.

3.2.2.2. Shannon Information, Communication and Control: The "Split" Between "Control" and "Why" Information

In order to show how information as a cognitivist concept lends itself to a problem-solving view of cognition, in this section, my goal is to elucidate this "split" between Shannon's quantitative conception of information on the one side, and, on the other, what in this section I will be calling "why" information. This "why" information refers to the more qualitative conception of information in which information is associated with meaning of the sort that is sought when we ask "why is this problem important" or "why should I have this goal" or "why should I care" – "what is the meaning of this task to me?". This kind of "why" information is part of what is taken as already-given in the cognitivist model because the problem with which the agent is contending constitutes a situation which the model takes as already-given. That the problem is of value to solve, a matter which is brought into question when we ask why, this is already-given in the model. In this way, this kind of "why" information is not something with which the modelled agent has to work or process. If the problem and its significance are already-given in this way, then the model leaves space only for problem-solving.

It is this point which I draw out in this section, arguing that cognitivist models of cognition work primarily with Shannon's conception of information, and that this conception leads to a problem-solving view of cognition. Information in this view is understood as a quantitative measure of a signal – a "message" – reproduced between two points. Core assumptions of cognitivist models of cognition are already present in this conception in the form of the passive, already-givenness of the information-processing agent and the situation they are in. Whilst my claim is not that this conception of information *caused* the problem-solving view distinctive of cognitivism, it is that this conception leaves space for little else.

To show this in detail, I point back to the cybernetic heritage of cognitivist mind research in which this problem-solving view made sense. Researchers at the time were trying to build engineering and electrical engineering systems – "control systems" (Wiener 1948/2013) – which had exogenously, pre-specified, that is, "already-given", goals (Gleick 2012, Dupuy 2000). Models for the workings of those systems, eventually described in computational language, became models for cognition, with the problem-solving framing imported in the background.

In the process of this history, the aforementioned "split" between the quantitative and qualitative senses of information became entangled and blurred. In this section I use the terms "control" and "why" information to distinguish between the two conceptions of information on either side of the split.

Following (Sloman 2014), I use "control" in reference to Norbert Wiener's (1948/2013) seminal book *Cybernetics or, Control and Communication in the Animal and the Machine*. One of the machines for which Wiener is famous for having worked on was an anti-aircraft weapons system. Such a system is built with a goal in mind, namely, shoot down aircraft. Again, the "control and communication" in question concerns the *internal coordination* of the weapons system. In the same way that the limbs of our body must be collectively coordinated in order for us to move with singular, coordinated purpose, the parts of the machine must similarly be coordinated as a whole. As each of the parts moves, coordination requires controlling or regulating the movement or behaviour of those parts relative to the whole, much as any coordinated human activity involves a certain amount of coordinated control over our body parts. This in turn requires accurate "communication" between the parts to remain coordinated[25]. If the communication through my body is such that my arm is receiving one signal and my foot another, it is going to be difficult to make progress toward my goal, (assuming the goal involves the use of my body). As we will see with Shannon's conception of information, communication requires an at-least-approximate fidelity in the signal. However, this sense of communication is speaking about the *internal* coordination of a system, not the more colloquial sense of, say, conversation and information-exchange between two human agents about anything in particular.

This is to say, Shannon's conception of information served to enable the modelling and construction of machines with pre-defined goals in which the engineering challenge was to ensure the machine could regulate itself through feedback loops in order to achieve the goal of its design. I use "control" information therefore to refer to information understood in this way.

To bring out the contrast with "control" information, I use "why" information to provoke an intuitive grasp of the other kind of information. The machines operating with "control" information operate with a certain independence relative to their goal because the goal is already-given to them. They do not care about the goal and "blindly" execute their function. Theirs is not to reason why, but simply solve the problem given to them. Integrating the discussion which on Superintelligence and existential risk which began the thesis, we can say that the goal is "orthogonal" to the machine's capacity achieve it.

I use "why" information therefore to identify the sense of information in which it conveys meaning. As I have been using the notion of meaning in this thesis, the meaning of something, for an agent, is the endogenously generated significance it takes on for that agent. The purpose of "why" information here is to pick out that kind of information in which, for example, an agent might ask why they should pursue a given goal, and a response will involve talking about or providing information concerning the meaning of that goal.

This, then, marks the split between "control" and "why" information. We can now turn to details concerning this split, beginning with Shannon (control) information.

It is well recognised and regularly emphasised in accounts of Shannon information, including his own accounts (1948, Shannon and Weaver 1949), that *information is not meaning* (Shannon 1948: 1, Dretske 1983: 57, Adams 2003: 474, Hidalgo 2016: xvi, DeDeo 2018: 12, Adriaans 2020).

[25] The earliest work in the Macy Conferences that were the provenance of the Cybernetic movement was focused on "feedback" in systems of different kinds (Wiener 1948/2013, Gleick 2012, Dupuy 2000, Rosenblueth et al. 1943). Goal-oriented systems must consistently regulate the behaviour of their parts, adjusting based on the feedback "communicated" through the system.

Instead, Shannon is concerned with the *quantity* of information in message. This is the "split". Information in terms of its meaning on the one hand, (its particular significance for an agent, from their perspective, and in a particular situation), and information as a quantity on the other. The way this split is usually characterised in philosophy is as a difference between "semantic" (meaning) and "non-semantic" (quantity) information (Floridi 2005, 2011; Piccinini and Scarantino 2010). Recall the discussion in the previous section on computation about the semantic account of computation and its difference to the others.

However, the semantic – non-semantic distinction is not as helpful for the purposes of understanding the problem-solving character of cognitivism. Aaron Sloman's (2014) distinction between information for "control" (non-semantic) and "communication" (semantic) is a more informative and explanatory account in this regard. It is also helpful for making sense of the distinction between the kinds of information machines seem capable of working with and the kind they do not – the kinds of difference that make a difference to machines, and the kinds that don't.

This claim about Sloman's control-communication distinction being more appropriate for distinguishing Shannon information from everyday senses of information is a working stipulation and assumption for this section. Sloman's characterisation helps to bring out features which resonate with the claims of this chapter as a whole. At the scale of granularity of a characterisation of the cognitivist paradigm, Sloman's distinction of information is more useful.

Whilst the semantic – non-semantic distinction tells us *what* the distinction is, Sloman's offers an explanation for *why* information-processing machines and computational systems work well with certain a kind of information (quantitative), but not another (qualitative), namely, because as *control* systems, their "why" is, again, already-given by the designer, builder, or modeller, and is thus not something with which they cognitively need to contend.

The processing of Shannon information concerns the coordination and control of a system in order to execute functions, solve problems, achieve goals and so on. Qualitative information involving understandings of why those functions, problems, or goals are valuable or meaningful, is not the kind of information that said system processes. Such engineering systems process information which effect functional changes ("stop", "go", "open switch") whilst not "understanding" those changes[26].

This is important to recognise because, in Shannon's time, Norbert Wiener had just published his seminal 1948 book, *Cybernetics, or Control and Communication in the Animal and the Machine* and the title can be misleading. Whilst Shannon explicitly says he is not concerned with the *meaning* of information, in the qualitative sense, he speaks nonetheless in terms of "communication", a term which has otherwise epistemic connotations and might suggest that a machine does after all understand what it is processing. Again, the "control" concerns only the coherent regulation and coordination of the limbs of, say, a robotic body; meanwhile "communication" refers to the internal passage of signals within that body necessary for the coordinated control. Finally, note in all this that for both terms "control" and "communication" to be used this way, the agent and their problem or goal is something that has to be already-given.

Continuing, Sloman's distinction is important then for making clear how this confusion can be misleading.

Again, what we get when we use Sloman's terminology is a clear sense that what Shannon was concerned with was a quantitative, non-semantic sense of information used in *control* systems, and *not*

[26] As Daniel Dennett's quip goes, there is "competence without comprehension".

the communication of meaningful, semantic information. Norbert Wiener, a colleague of Shannon's, who will be discussed later in this section for his work on cybernetics, spoke of "communication *engineering*" – that the focus was on engineering systems, rather than a theory of mind or cognition, is important to keep in mind (Wiener 2013: 10).

In any case, it is important to note that a lot of theoretical mileage is made in cognitivist, information-processing accounts of mind, under the cover of this confusion and ambiguity. In what follows I will try to be as explicit as possible.

The split between "control" and "communication" information, in Sloman's sense, is best understood if Shannon information is introduced first, so after an account of Shannon information in everyday language, I will say more about the notion of "control". Here is an oft-quoted passage from (Shannon 1948):

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. *The significant aspect is that the actual message is one selected from a set of possible messages*. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. (Shannon 1948: 1) (emphasis my own)

The "communication" in Shannon's sense here is the reproduction at point B of a "message" sent from point A. For Shannon, building on Nyquist (1924) and Hartley (1928) before him, the concern was principally electronic "communication systems" which, at the time, were "systems of telegraphy, telephony, picture transmission and television over both wire and radio paths" (Hartley 1928: 535). All three men were engineers at Bell Labs, a company based in New Jersey and then New York with a historic blend of scientific and engineering research and development (Gleick 2012: 188-189). All the men were concerned with optimising the performance of electrical communications system and so were researching and formalising the capacities of electrical wires, (connecting an exploding network of telephones), to "reproduce" signals at either end. Bell Labs was set up by the Alexander Graham Bell, inventor of the telephone, with money from the Volta Prize, awarded to him (Bell) by the French Government in 1880. Having followed fairly swiftly from the development of the telegraph, the telephone:

> "…had to be explained, and generally this began by comparison to telegraphy. There were a transmitter and receiver, and wires connected them, and *something* was carried along the wire in the form of electricity. In the case of the telephone, that thing was sound, simply converted from waves of pressure in the air to waves of electric current." (Gleick 2012: 190) (emphasis in original)

In this case, the concern of Shannon's "communication" is to make sure that the spoken "message" is accurately reproduced on the other end as the signal is transformed between air pressure and electric current. From Shannon's engineering perspective, the semantic content of the message is irrelevant to this problem of ensuring that the sent signal is reproduced – "communicated" – at the second point. He showed that it is possible to do this with a purely quantitative conception of information.

Again, Shannon pulled directly from Nyquist's (1924) and Hartley's earlier (1928) research, available at Bell Labs. In this quantitative conception, *information is an expression of choice amongst possible alternatives*. "The significant aspect is that the actual message is one selected from a set of possible messages." (Shannon 1948: 1) Hartley gives as example the "message" that "Apples are red", noting that:

> "…the first word eliminates all other kinds of fruit and all other objects in general. The second directs attention to some property or condition or apples, and the third eliminates other possible colours. It does not, however, eliminate possibilities regarding the size of apples, and this further information may be conveyed by subsequent selections."(Hartley 1928: 536)

Each word is a selection from a set of possible words, different differences. A three-word sentence is not a long one, but in the English language the set of possible words is enormous and so the selection of each word is one from a large set. Quantitative, non-semantic, "control" information is the quantity of this proportion in a selection. Hartley and Shannon were not linguists looking for laws of language, just creative engineers trying to optimise the performance of electrical communication systems. Energy and cost could be saved if the same information could be coded into a smaller message. They recognised that, given a first word, there are certain words which are more and less likely to follow. In the English language, "are" is a highly likely word to follow a plural noun like "apples". Rather than each word being a selection from the same enormous original set then, Shannon also recognised that each word or symbol affected the probability of that which followed. We recognise that, for example, after a "q", a "u" is very likely, and that a consonant is very *un*likely after an "an", that certain words are more likely after the word "yellow", and lots of words are *less* likely after the word "information", and so on. Shannon information is often spoken of as a measure of "surprise" because the next letter or word can be more or less of a surprise (Gleick 2012: 216).

"Surprise" and information are therefore complimentary views on the same thing in this context – the more surprise, the more information. The defining moment of the "three men walk into a bar" kind of joke is the punchline. There is little "surprise" in the second man who repeats the action of the first, and lots of humorous surprise in the third who breaks the pattern. To be clear, in this case, it is not that the third man contains more surprise or information than the others. The joke is only amounts to something funny or surprising as a whole. Taken individually, the actions of each man are not funny or surprising. The sequence of the whole is necessary, like with notes in a melody. No note has more information than any other, but put together in sequence and rhythm, the whole becomes something interesting.

This surprise-information relationship was an important insight for the engineering problem because it meant that a lot of redundancy could be jettisoned, saving energy and money in the (quantitative) "communication" of messages. Because there is very little surprise in "u" following "q" in the English language, the "u" offers very little information. This "redundancy" was important. Shannon represented

it with the variable *D*. Gleick quotes Shannon: "*D* measures, in a sense, how much a text in the language can be reduced in length without losing any information." (Gleick 2012: 216)

Now, as far as the "choice from a set of possible alternatives is concerned", the simplest choice is binary – a choice of one from a set of two. This minimal amount of information is the famous "bit" and is often explained in terms of a coin flip, a single outcome, or "choice", from a set of two. For his master's thesis, Shannon showed how Boolean logic could be applied to electric circuits. At the heart of Boolean logic is the idea that all values are a "choice" from a set of two: truth or falsity. Now, unlike letters and words, and taken in isolation, a single coin flip does not make a difference to the likelihood of the outcome of any another flip. Like letters and words though, each choice is discrete, but Shannon also worked out how to quantify information for continuous phenomena. In his (1948) paper, Shannon speaks of both discrete and continuous information. It is much simpler to imagine a "choice from a set of possible alternatives" by working with discrete things like strings of letters or 0's and 1's, but of course the media with which Shannon and others were working - sound waves, radio waves, electric currents and the like – are continuous phenomena. Gleick notes Nyquist's (1924) simple solution:

> "It had been known since the dawn of telegraphy that the fundamental units of messaging were discrete: dots and dashes. It became equally obvious in the telephone era that, on the contrary, useful information was continuous: sounds and colours, shading into one another, blending seamlessly along a spectrum of frequencies…Nowadays most of the current in a telegraph line was being wasted...it was a case of amplitude modulation, in which the only interesting amplitudes were *on* and *off*. By treating the telegraph signals as pulses in the shape of waveforms, engineers could speed their transmission and could combine them in a single circuit – could combine them, too, with voice channels…Nyquist's method was to sample the waves at intervals, in effect converting them into countable pieces." (Gleick 2012: 199) (emphasis in original)

Working from such methods, the information in a transmission can be calculated in terms of a series of choices from a set of possible alternatives. Shannon took a statistical approach to the matter, formalising his theory in terms of probability – what is the likelihood of a given letter, say, given what has already come in the message. The more likely, the less information (recall "u" after "q"). The less likely, the more information (the third man in the bar). Hartley (1928) has a simpler formula expressing the same basics as Shannon:

$$H = n \log s$$

H is the amount of information, *n* is the number of symbols transmitted in the message, and *s* is the size of the alphabet from which each symbol is a choice. As far as the logarithmic weighting goes, Shannon agreed with Hartley that it was simpler for several reasons, including simply that:

> "It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time

> roughly squares the number of possible messages, or doubles the logarithm, etc." (Shannon 1948: 1)

As a mathematical tool, logarithms reduce more complicated multiplications (e.g. the doublings which create exponential increases) to simple additions.

Building on Hartley, Shannon recognised that each sequential "choice" of symbol will have a probability, given the available choices, and the "history", so to speak, of what has come before it. "A message, as Shannon saw, can behave like a dynamical system whose future is conditioned by its past history" (Gleick 2012: 226). These probabilities can be taken together as a sum, and that is how Shannon formalised it, information as the sum of probabilities of each of the parts of the message.

$$H = -\sum p_i \log_2 p_i$$

Here p_i is the probability of each message and *H* is what has come to be known as "the "entropy" of a message, or the Shannon entropy, or, simply, the information." (Gleick 2012: 229). The negative sum may be confusing, and it takes some minor and unnecessary-for-purposes-here detail to explain. Shannon's negative formulation is just one way of framing the complementarity, between information-as-surprise and entropy. Information for Shannon involves many entangled notions - likelihood, probability, "surprise", "unexpectedness", "entropy", and also "uncertainty". For him it is a solution to the question: "Can we find a measure of how much 'choice' is involved in the selection of the event or of how uncertain we are of the outcome."[27] (Shannon 1948: 10)

It is important to say more about this connection with entropy. It leads to Sloman's distinction between (quantitative, non-semantic) "control" information and (semantic) "communication"[28].

The link between information and entropy is clearest in James Clerk Maxwell's famous thought experiment, "Maxwell's Demon". A little history goes a long way here. These ideas and the conception of mind they midwifed all seemed to be circling around the same mysteries. Maxwell presented his thought experiment officially in 1872 in his book *Theory of Heat*, where he used the term "finite being" to describe the "demon"; the term "demon" was coined two years later by William Thomson (1874), aka Lord Kelvin, in his paper "The Kinetic Theory of Dissipation" (Gleick 2012). The thought experiment brings together many central ideas relevant to an understanding of information – entropy, dis/order, uncertainty, probability, and, importantly, observer effects. Maxwell and others were developing the science of thermodynamics, the dynamics of heat, i.e., the energy/movement of particles, particles whose existence would come to be proven by Einstein in his 1905 paper on Brownian motion. Though the language of statistical physics did not emerge until after Einstein's paper, Maxwell still understood thermodynamics in a highly probabilistic way:

[27] Note that what is of interest to Shannon is, in the end, a "measure", a distinctly quantitative phenomenon.
[28] Recall that Shannon used the term "communication" but was concerned, again, with quantitative and non-semantic information. Following Sloman, this is what I have been referring to as "control" information.

> "The 2nd law of Thermodynamics has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again." (Gleick 2012: 274)

There is a directionality, and an irreversibility to the direction. A "certain" "expectedness". Nothing in principle denies the possibility of getting the same tumblerful of water out again, it is just extremely unlikely because the likelihood of the tumblerful of water remaining ordered in the way that it was in the tumbler is extremely low once it interacts with the sea. Gleick puts it like this: "The second law, then, is the tendency of the universe to flow from less likely (orderly) to more likely (disorderly) macrostates." (Gleick 2012: 275). The important thing is that it takes work to create order. (The notion of work will be explored in detail in section 4.3.2.2 in the context of "thermodynamic work cycles", in a discussion on living systems.) If things are left to themselves, entropy increases. The usual addendum is that this is the case for "closed" systems, (e.g., a gas canister), but not for "open" systems (e.g. hydrological systems).

"Maxwell's Demon" therefore bridges a probabilistic understanding of entropy with information. The thought experiment involves two chambers of gas separated by a (massless) door, a door operated by a demon. We are invited to imagine that the demon can perceive which of the gas particles are fast and which are slow. When the particles are mixed, there is a high state of entropy or disorder as the gas moves toward an equilibrium. As they are bouncing around, the demon selectively raises the door to allow the fast molecules to pass to one side, and the slow molecules to the other, creating a chamber of hotter (faster) gas, and a chamber of cooler (slower) gas. The conundrum the thought experiment presented at the time was that the Demon did no work on the gases, and yet the gases moved into a more orderly state. The demon thus defies the probabilities of entropy.

Interesting in its own right, but also for its (retrospective) insight into the cognitivist paradigm, a recognised "solution" was presented by a Hungarian physicist Leo Szilàrd in 1929 (Gleick 2012: 279). "The demon replaces chance with purpose. It uses information to reduce entropy." (ibid: 276). The idea was that the information the Demon observed about the particles, whether they were fast or slow, was not free. It cost something. It takes energy to get the information about the particle. The *observation of the system is involved with the system*, expressing thermodynamic work on it, which makes a difference to the system, enabling it to run counter to entropic probabilities[29].

Szilard had previously been working on perpetual motion machines, and the influence in thinking was apparent (Gleick 2012: 279). His "solution" to Maxwell's demon is the same reason perpetual motion machines are not possible, or, at least, in both cases there is a bringing of awareness to the way something thought to be outside the system must actually participate in the system in order for it to work. In the case of a perpetual motion machine, it is now well recognised that work expresses heat (energy) which is lost from the system, and that unless that energy is resupplied from something outside the system, said system will eventually lose all its energy and come to a stop. The machine cannot fully capture the heat it generates. It would require a second machine to capture that heat. That second machine could be integrated with the first such that they were one system, but that second part would itself also generate heat which would require a third to capture, and so on.

[29] Thinking about how "observation" is understood here, later in this chapter I will talk about the influence of vision neuroscientist David Marr in the cognitivist paradigm (§3.2.3.4.), and in particular, the biases present when observation-as-vision is the metaphor for the cognitive process, for the way in which vision affords a sense of being *outside* the system in question, as the Demon was originally imagined to be.

The significance of this thought experiment here suggests that a fuller account of the cognitivist paradigm would greatly benefit from a history of philosophy perspective because there are many contemporaneous developments in philosophy that were entangled. Perhaps most famously, Kurt Gödel's work on the Incompleteness of formal systems was published in 1931, just two years after Szilard's proposal to Maxwell's Demon. Both express the point that a thing cannot fully contain itself (without inconsistency) and, more subtly, they both invite an awareness to (participation in) a larger system transcendentally required for operation. Another entangled and contemporaneous development[30], whose connection to the historical development of models of mind deserves further exploration, was the respective interpretations of quantum mechanical phenomena by Niels Bohr and Werner Heisenberg – in at least in one interpretation, they differed on what counted as "inside" and "outside" the arrangement experimental apparatus necessary to produce the observations (Barad 2007).

There is in these examples an expression of a core assumption about cognition in the cognitivist paradigm. Again, it is the idea that we are "outside" the thing we are coming to know. Or, that we do not influence it. In Shannon's model of "communication" this makes sense though. The sender and receiver do not influence the message or its probabilities, these simply send or receive. Further, both sender and receiver, and message, are assumed. These "two points" between which the message is to be "reproduced" are "already-givens". In his model, Shannon takes up a position outside this theoretical dynamic much as he physically stands "outside" any electric circuit with which he might be tinkering. His observation of the system when it executes does not interfere with it in the way that Maxwell's demon does. This is to say, it stands to reason that a model of mind informed by such a conception of information-processing would generate an account of cognition as passive perception of a pre-defined "message".

And yet, what Szilard showed is that *information is not free*. It is not just abstract. Processing information is an interaction which costs something in the form of physical entropy. As such it is a form of participation, like Maxwell's demon, effecting the entropic state of the order of the whole system.

According to Gleick:

> "Szilàrd had thus closed a loop leading to Shannon's conception of entropy as information… To the physicist, entropy is a measure of uncertainty about the *state of a physical system*: one state among all the possible states it can be in. These microstates may not be equally likely, so the physicist writes $S = -\sum p_i \, log \, p_i$
>
> To the information theorist, entropy is a measure of uncertainty about a *message*: one message among all the possible messages that a communications source can produce. The possible messages may not be equally likely, so Shannon wrote $= -\sum p_i \, log \, p_i$.
>
> It is not just coincidence of formalism: nature providing similar answers to similar problems. It is all one problem. To reduce entropy in a box of gas, to perform useful work, one pays a price in information. Likewise, a particular message reduces the entropy in the ensemble of possible messages – in terms of dynamical systems, a phase space." (Gleick 2012: 280) (emphasis my own)

[30] Yet another historical trajectory of philosophical ideas which merits further exploration is the role and "initial condition" played by the logical positivism/empiricism of the Vienna Circle. In his 1948 book introducing Cybernetics, Norbert Wiener acknowledges the heritage of mathematical logic, the language espoused by most of the Circle, even choosing Leibniz as "patron saint" (Wiener 1948/2013: 12)

Entropy and information are the necessarily complimentary sides of that sheet of paper down which a line is drawn, generating both sides. Now, finally, the line from entropy and information to Sloman's characterisation of the split between "control" and "communication" can be made. The line leads through cybernetics, a field of research developed by Norbert Weiner (2013).

Norbert Weiner, another Bell labs man, had very similar ideas to Shannon, though he (Wiener) followed them in different directions. Being abreast of the developing work in statistical mechanics in physics, Wiener shared Shannon's statistical and probabilistic approach to information as an approach to calculating the likelihood of a "choice amongst possible differences":

> "We shall see that this dominance of statistical mechanics in modern physics has a very vital significance for the interpretation of the nature of time. In the case of communication engineering, however, the significance of the statistical element is immediately apparent. The transmission of information is impossible save as a transmission of alternatives." (Wiener 2013: 10)

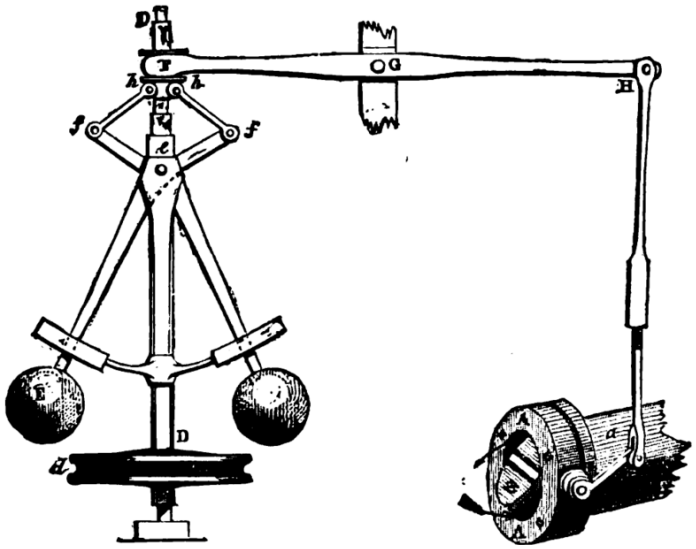
Shannon and Wiener were on sides of the sheet of paper. "For Wiener, entropy was a measure of disorder; for Shannon, of uncertainty. Fundamentally, as they were realizing, these were the same." Gleick 2012: 247). As Wiener puts it:

> "The notion of the amount of information attaches itself very naturally to a classical notion in statistical mechanics: that of *entropy*. Just as the amount of information in a system is a measure of its degree of organization, so the entropy of a system is a measure of its degree of disorganization; and the one is simply the negative of the other. This point of view leads us to a number of considerations concerning the second law of thermodynamics, and to a study of the possibility of the so-called Maxwell demons." (Wiener 2013: 11) (emphasis in original)

It is from this conception of information that Wiener moves in establishing the field of cybernetics. It is in the field of cybernetics that the conception of Shannon information in Sloman's term of "control" is most clear. Wiener makes the conceptual link between what he sees as "the essential unity of the set of problems centering about communication, control, and statistical mechanics, whether in the machine or in living tissues." (Wiener 2013: 11) The conceptual link he finds, at least in "communication and control", concerns *feedback*. In dubbing the new field "cybernetics" – from "steersman", "governor"[31] in Greek – he acknowledges the "the first significant paper on feedback mechanisms is an article on governors, published by Clerk Maxwell in 1868…We also wish to refer to the fact that the steering engines of a ship are indeed on of the earliest and best-developed forms of feedback mechanisms." (Wiener 2013: 11-12)

[31] The connection seems to be that a steersman guides or controls his ship through the variations of the sea, constantly adjusting in continuous feedback loops between his steering, and the effects of waves and currents and winds on the ship.

"Governor" and "steersman" quickly lend themselves to a connection with the notion of control. From Maxwell's (1868) discussion "On governors"[32], Wiener took the power of feedback as a control mechanism even further. A governor in the sense of the term here, as opposed to the political station, is a small feedback system used to control the operation of a machine in some capacity. It was used by James Watt on a steam engine even before Maxwell's paper and continues to be used today in cruise control systems in cars. In older systems, like the "centrifugal governor" below, two spinning arms with balls on the end were connected to a steam valve (right of image). There was a basic feedback between the opening of the valve and the speed of the spinning balls whereby, if the valve opened, the balls would spin faster, but as they spun faster, the arms would raise, which would pull down the short levers above, pulling down the long lever connecting the centrifuge to the valve, which in turn closed the valve, slowing the centrifuge, and reopening the valve, and so on, creating a "dynamic stability" (Maxwell 1868).



3.2.2.3. Control Information: Definition

For the purposes of precision, we can offer a more compressed definition now of control information:

> A signal is "control information" if and only if it is "communicated" between points of a system in a way that serves the control and coordination of that system, independent of the purpose of that control and coordination.

In simpler terms, we might call this "tell me what to do" information for the way in which the kind of information in being told what to do is orthogonal to the kind of information in being told "why" that thing is worth doing. "Tell me what to do" or "control" information does not include *why* that thing needs doing from the perspective of an agent, or why, for and by means endogenous to the agent, they

[32] See (Mayr 1971) for a more recent detailed discussion of (Maxwell 1868).

encounter it as significant. Control information does not need to include this. All that is necessary for the controlled and coordinated functioning of an agent, from the external perspective that an engineer such as Shannon takes on, is the "tell me what to do" kind of information. A machine does not need to know why it is doing what is doing in order to do what it is supposed to do. This "why" information which explains the meaning of the task, may "inform" understanding, but it does not affect action in the same way. An anti-aircraft weapons system does not need this information to do what it is designed to do. Such information is, again, "orthogonal" to the information necessary for coordinated functioning. The "functioning" is an important part to highlight, for there is an inherent functionalism deeply rooted in this Shannon/"control" conception of information and this functionalism is one of the major characteristics of cognitivism. More will be said about functionalism in the next section.

There is a subtle but important observation to make about control information following the definition above. It can be approached by noticing the source of meaning of the final goal or function. The control information being "communicated" between points of an anti-aircraft weapons system does not ask for "why" information. The "why" is specified *externally*, by the agent/human who designs, builds and models the system, (and the operator who fires it in the particular direction).

In the case of a model of cognition, this is the equivalent of the modeller exogenously projecting some meaning of the modelled agent's cognition onto that agent. Whilst the projection may be more or less aligned, this means what is significant for the machine or agent is assumed exogenously by an agent with a 3rd person perspective on the machine. In building the system in the first place, designing the network of communication channels between the "points" and parts of the system, what is to count as signal and noise for the system is also determined from this outside, 3rd person position.

If, for example, another human provides me with control information, "communicating" what to do, I want to know why it is important to do what it is I am being asked to do. Whilst control information is about a "choice amongst possible alternatives", to understand the meaning or "mattering" of task, I want to know about the other possible choices and why, given the "possible alternatives", this particular choice was chosen. Ideally, I also want some agency in determining the choice amongst possible alternatives. Without this aspect, the particular significance of the choice to me is not mine to realise. Instead, it is received as an order - information "communicated" for the purposes of "control" and "coordination", and the significance is decided by somebody else. What matters is not my concern. Tennyson's famous lines from his poem "The Charge of the Light Brigade" come to mind: "theirs not to reply/ theirs not to reason why/ theirs but to do and die". Hierarchical systems of organisation, like military and corporate organisations "function" to an important degree with "communication" of "control" information, but not "why" information.

To wrap up this section on control information, I want to be explicit about how it contributes to a problem-solving conception of cognition. The starting position is that information has, in the philosophy of mind and AI, at least two senses. Shannon's is the more technical and developed account, the one I have been referring to here as "control" information. The other is the more everyday sense of information informing me about something or conveying meaning for me. I have been pointing to this other sense with "why" information. "Why" information should not be taken as a complete account of meaning or anything like that. The purpose it serves to point to a basic distinction. If I design and build a robot to open the fridge and bring me yesterday's leftovers, the robot does not need to care about that goal. It does not need "why" information about what it is doing in that respect. What is does need, is information to be accurately "communicated" through its body so that it is sufficiently coordinated as an entity to perform the relevant actions for achieving that goal. This is the sense of information that is useful to the robot and to the designer and engineer thereof. The key, then, is that the goal is taken for granted, already-given. Thus, when it comes to using this situation or any analogous situation with AI

as a model for the cognitive process, as cognitivism does, *the model focuses on the way the entity solves the already-given problem.*

More precisely, "control" information is the sense of information used to model and build systems that solve already-given problems. Systems of robotics and AI like these stand as the cognitivist models of the cognitivist process, and lo, cognitivism models cognition as a process of problem-solving.

Before moving on to talk about functionalism as the third main cognitivist concept, I want to finish this discussion of the place of information in cognitivism by noting where it shows up in Western philosophy of mind. In the same way that it supports a problem-solving view of cognition, I want to show how it can be traced a source of some classic problems in philosophy of mind. In noticing such connections, it becomes possible to notice an ecosystem of concepts and thought which is self-supporting. In the end, my goal in this thesis is to show that cognitivism and post-cognitivism are two different ecosystems, not irreconcilable – a matter beyond this thesis – and as such, produce different creatures, concepts, and ways of thinking about AI.

3.2.2.4. Connecting Shannon Information to Classic Problems in Western Philosophy of Mind: The Chinese Room and the Imitation Game

Reiterating the conversation above, just because a signal is reproduced with a certain level of fidelity at two points across a medium, doesn't mean it is "communication" in an everyday sense of communication between agents. The "control" conception of information refers to a quantitative measurement in Shannon's conception as "a choice amongst possible alternatives". Heads and tails are each one choice out of two alternative possibilities, a letter in the Latin alphabet is one out of 26 and therefore constitutes "more" information, and our planet is one amongst quite a lot more than that. The question of information in this context is "how much" information, measured, again, in terms of how unlikely the selection is, how much of a "surprise"[33] it is, given the available alternatives. As various authors in this context point out (Shannon 1948: 1, Dretske 1983: 57, Adams 2003: 474, Hidalgo 2016: xvi, DeDeo 2018: 12, Adriaans 2020), information is not meaning, a term I have been defining as the significance an agent endogenously encounters. The key, again, for this section on information, is that, this "control" sense of information leads to, or supports[34] a problem-solving conception of cognition. Now I want to point to two expressions or symptoms of this conceptual ecosystems that show up in contemporary Western philosophy of mind.

I consider John Searle's Chinese Room thought experiment first. My intent is not to "solve" it in any way but to point out that the construction of the thought experiment assumes the same already-givens distinctive of a cognitivist model of cognition and turns on the distinction between "control" and "why" information even if it is not articulated in those terms. I then also consider the Imitation Game, otherwise known as the Turing Test, which makes the same assumptions and moves, showing that information in

[33] Return to the discussion on page 72 for a reminder on "surprise" as it pertains to information.

[34] Again, it is tempting to imagine that the development of information theory and the "control" interpretation of information causally *led* in some way to a particular view of cognition as information theory increasingly became a feature in the cognitive sciences after the Macy Conferences (see Dupuy 2000). The precise nature of the influence of these concepts is historically and culturally more complex than a sequential causal connection and I do not know the history well enough to comment. What I do wish to say here is that the concepts fit together and mutually suppose each other such that, if one adopts a control conception of information, one then has to adopt a problem-solving conception of cognition too, and vice versa. My claim is that this mutual supposing of each other is rooted in their shared assumption, namely, taking the cognitive situation as already-given.

Shannon's sense has been influencing how cognition is framed and modelled in the cognitivist paradigm since its inception.

In John Searle's "Chinese Room" thought experiment, the reader is presented with a situation in which a human is successfully – functionally – processing symbols without any comprehension of the symbols (Searle 1980). Given some "input" of Chinese symbols into the room, the human has a book which specifies what symbols to respond with. Exchange happens which we would intuitively call "communication", not in Shannon's sense of isomorphic reproduction between two points, but just conversation, dialogue etc. The puzzle is that the human in question understands none of it. None of it makes a difference to him as anything more than "control" information specifying what actions to take. That is, as per Sloman's note on control information above, the symbols processed by the man effect and modify his actions, making a difference *in the "control" sense*. However, they do not communicate "why" meaning to him – there is no particular significance for him, and from his perspective to the particular transformations he is effecting. They make no difference to him *in the "why" sense*. The meaning of what was "communicated" is exogenous to the human in the thought experiment.

This line of thinking also invites an interesting response to Turing's question "can machines think" that he raises in "Computing Machinery and Intelligence" (1950). In that paper he introduces the now in/famous "imitation game".

The important point here is that nature of the game does not enable us to really distinguish between control and communication, between a machine doing what it is told, and having a sense of the meaning of things. With the appropriate program, the machine in Turing's imitation game can do just what Searle's human does in Searle's thought experiment. The information coming in can effect modifications in the output of the machine and, much as with Searle's human, we are to determine whether there is thinking (Turing's term) or "intentionality" (Searle's) from this output. The problem is that the output, in both cases, can be produced with control information. In Searle's case it is the book of rules which specify particular mappings between symbols, and in Turing's case it will similarly be a program or algorithm.

The imitation game is supposed to be conducted by interface, as an exchange of digital text. The point of doing it this way is so that the obvious physical differences between humans and computing machinery are not differences used to determine which is which. The digital interface means the functional performance is the only difference that matters.

> The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a 'thinking machine' more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices…The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include. We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant. (Turing 1950: 434-435)

However, what Turing points to as "disabilities" are not arbitrary matters. The conditions of the game *do* seem to make a difference because they involve an elimination of variables such that the only difference-making differences that matter are those of which a computer is capable. By reducing to a lowest-common-denominator "question and answer method", Turing foreshadows the disembodied functionalism and "multiple realisability" that permeated thinking about mind in the cognitivism tradition and which I discuss in the next section on functionalism.

The notion of "control" information is useful therefore here for making sense of Turing and Searle's thought experiments because we can account for a successful "imitation" and "Chinese dialogue" by saying that in both cases there is "control but not communication"[35] - functional competence but not comprehension. In both cases, intelligent behaviour and even understanding can be imitated using only "control" information, and whilst Turing's example leaves it open for debate, the point of Searle's thought experiment is to show that this imitation is insufficient for understanding or any kind of sense of involvement of meaning.

With that in place, I can offer a brief summary of where we have come to. So far, I have characterised cognitivism in terms of its "problem-solving" orientation. I begin by presenting and defining computation, the first major concept of cognitivism, as rule-defined information processing. I noted the notion of computation requires taking the cognitive situation itself as already-given and in this way, supports a problem-solving view of cognition. We can now see that the notion of "information" fits very neatly into this picture. The notion of information at work is in fact the "control" sense of information, a sense which is distinct and orthogonal to what I have been describing as "why" meaning. In using the control sense of information as cognitivist and contemporary AI work does, the meaning of this information to any agent in question, has to be already-given in a model of that process. In short, in the same way that a model of cognition as computation naturally leads to a conception of mind as a problem-solving thing, "control" information naturally leads there too.

With the notions of computation and information in place, we can turn to functionalism, the last of the three concepts used to characterise cognitivism here.

3.2.3. Functionalism

> Turing-style functionalism constitutes the heart of what is called "cognitivism", which remains today the dominant paradigm of the sciences of cognition.
>
> - J-P. Dupuy, *The Mechanisation of the Mind*

Functionalism is the last leg in the fundamental conceptual structure of cognitivism such as I have articulated its joints in this thesis. Like computation and information, my work is to show that the concept is a feature in a model of mind which takes the very cognitive situation as already-given and, in this way, generates for cognitivism a problem-solving conception of cognition.

To support this position, I begin with a brief discussion of the way in which "function" is intuitively used in language to bring to attention the kind of assumptions that are involved even in our everyday use of the terms. From there I move into a more technical discussion of functionalism and offer a basic

[35] This is an echo again of Daniel Dennett's famous notion of "competence without comprehension" – see for e.g. Dennett's (2017) *From Bacteria to Bach and Back* for a thorough, book-length discussion.

intellectual history. I then present the fundamentals of the main accounts of functionalism in philosophy of mind, focusing on the accounts first articulated in Hillary Putnam (1960, 1967). From there I segue to Marr's famous Tri-Level Hypothesis (Marr 1982) which took functionalism from a philosophical view about mental states to a generalisable scientific methodology involving the decomposition of an information processing system into three main layers of analysis. From the perspective of this thesis, this dissection of information-processing systems into functional aggregates is a key expression of the fundamental place (and function) of functionalism in cognitivism. With this theoretical background in place, the way in which functionalism participates in cognitivism's problem-solving conception of cognition is the much clearer. Making that connection is the aim of this larger section on functionalism, the final section in this chapter on cognitivism.

3.2.3.1. A Quick Note on the Everyday Language

I want to make two small observations concerning the way in which the notion and word "function" is used in everyday language. One is the function/malfunction duality in which case "function" operates as a verb and we might ask "is the machine functioning", is it working, or is it doing the thing it is supposed to be doing. This "functioning" in turn presumes the second sense of the term - "function" as a noun, in which the "function" of a machine is the purpose, end, or goal of the machine, and with respect to which the function/malfunction distinction is made. The "function" of calculators is to solve problems of arithmetic. The function of cars is faster-than-human travel. The function of a hammer is to leverage greater-than-human force. Whilst it hardly counts as an argument, I nonetheless find the way we use the language like this, with the bias to specific goal-oriented problem-solving it seems to involve, something that should be kept in sight when reflecting on the model of mind which it supports. Notice also that the function in each of these cases, the thing each thing in each case is supposed to be doing, is determined exogenously, by an agency external to it.

Continuing this line of thought for a moment, of course each of the above objects can be used for more than one purpose or end. Recall from the previous section on information the rhetorical question "what is the function of a screwdriver". In that context the question was leveraged to make sense of Gregory Bateson's conception of information as "the difference that makes a difference". It was noted that the ways in which a screwdriver can make a difference are many, (doorstop, back-scratcher, conductor's baton, and so on), and that the "function" of a screwdriver is highly contextualised. Whatever the context though, the particular function or end for which a screwdriver is used, is not something the screwdriver itself determines – it is not an agent. The screwdriver is used as a means to an end not its own. The screwdriver has no say in whether it is used as a chair-leg, door-stop, or paper weight. The function of the screwdriver is therefore something exogenously specified.

There is a wealth of literature which goes into depth concerning the way in which we speak of functions, both when we speak of biological or evolutionary functions, as well as in when we speak of technological objects, or "artifacts". John Searle (1995) articulated the difference by speaking of "agentive" and "non-agentive" functions. The difference is usually taken to be a matter of the place of intentions in determining what the function in question is (Neander 1991, Houkes and Vermaas 2004,2010).

My concern is not to enter this discussion. I want instead to bring attention to those things which, as Francisco Varela, pioneer of the Enactive paradigm of mind research I discuss in the fifth chapter, put

it - "are always before our eyes" (Vörös 2023)[36]. Like computation and information, the discussion of functionalism brings with it certain presuppositions, specifically the already-given cognitive situation. The effect is that we can miss in models of mind and cognition that the function of mental states is something exogenously determined. In this way, functionalism propagates the logic of the cognitivist model, or the way has things set up.

This point concerns the methodology of the cognitivist model and the effects of the already-given order of things on how we are subsequently able to conceive of cognition. When the function in question is determined from an exogenous, 3rd person perspective, the focus of the model becomes to explain how that function is executed or realised in the modelled agent and does not account for the meaning of the function to the agent, (this is taken as already-given), let alone from the 1st person perspective of that agent.

This is not a problem for a model of cognition per se. In fact, "function" works very well as a concept for describing the functioning of objects *because* it excludes these things from the model of the object.

However, it should then be questioned whether this way of modelling is appropriate for modelling *human* minds because, insofar as human minds[37] are caring and discerning about what they do, a complete model of the human mind would at a minimum need to include this much. As it stands, these questions are *not* something with which cognitivism concerns itself, still less AI research which is a distilled expression of the cognitivist model.

As I now move to consider functionalism's early history, it will be useful to keep this in mind and notice the way it emerges when functionalism was introduced in philosophy of mind.

3.2.3.2. Functionalism in Analytic Philosophy of Mind

The first discussion of a functionalist approach in shows up in Putnam's "Minds and Machines" (1960) and, more explicitly, in "The Nature of Mental States" (1967). The questions motivating these papers are shared. Putnam was trying to figure out how make sense of identity statements or questions like "is pain a brain state?" As noted earlier in section 3.1.1. (Cognitivism in the Literature), the "cognitivist" revolution in psychology in the 1950's was responding to the Behaviourism of the time (Wallace et al. 2007, Varela et al. 1991: 6, Thompson 2007: 4-5). In the eyes of behaviourism, questions about interior mental states like pain were not permitted because "brain state" refers to something internal, and thus metaphysically speculative and poorly defined. Instead, Behaviourism endeavoured to sufficiently explain "pain", for example, in terms of clearly defined behaviours. However, in Chomsky's (1959) review of Skinner's (1957) *Verbal Behavior*, Chomsky makes a now-recognised critique, pointing out that behaviourism depends on *implicit* reference and control of variables to do with internal states anyway. For example, explaining the behaviour of a lab rat requires statements to do with the rat feeling hunger and wanting food. The ideas of "computer", "computation" and "information-processing" permitted reference to internal states and processes, like "brain states", that explained such behaviours

[36] As Vörös notes, Varela was in fact quoting a line from Wittgenstein of which he was fond: *"*One is unable to notice something — because it is always before one's eyes" from (Wittgenstein 2001: §129).

[37] A post-cognitivist might want to broaden beyond just human to include living systems more generally, with a line somewhere and subject to debate, but I am focusing on human minds for the moment here for the sake of the argument.

in more acceptable terms. A major part of that was to do with the functionalist terms in which it was possible to then explain behaviour. Functionalism did not involve anything mysterious.

From the perspective of this thesis and a post-cognitivist position, functionalism took this ground because it is in fact not so different from behaviourism. It is simply a behaviourism about "internal" states. The naissance of cognitive psychology chunked the mind into functional modules, each explaining the behaviour of a human by reference to the behaviour of a functionalist taxonomy of modules and mechanisms like "attention", "memory", a kind of thinking Jerry Fodor brought into Western philosophy of mind with his (1983) *The Modularity of Mind.* There is a history of discussion about whether this modularity is just the case for "lower-level" functions to do with basic perception, or whether it includes "higher-level" modules like "reasoning" and "planning". See Sperber (2002) and Carruthers (2006) for "post-Fodorian" discussions of modularity of mind. There is no need to go into more detail than this, the point is that functionalism is usually dialectically taken as a response to behaviourism, but functionalism carries a behaviourism with it.

So, as part of his work on the mind-body problem, Putnam was trying to make sense of propositions like "is pain a brain state" in a way that moved beyond behaviourism, to see "whether or not it is ever permissible to identify mental states and physical events." (1960: 20) In the course of this paper, he presented what has come to be known as "machine state functionalism", the first of three different kinds of functionalism that I will discuss.

In the history of philosophy of mind, functionalism arrived on the scene to offer an answer to the same question that philosophers sought to answer concerning computationalism, namely, how do we identify and individuate mental or cognitive[38] processes[39] as such. Recall there was an intuition that not all transformations which might fit the bill actually count as computation and so the question arose of how to distinguish what are and are not computations in those transformations which we observe in systems of different kinds. Harking back to the discussion of computationalism, the inquiry took the particular for in the form of the question, "when is a physical system computing?" Would we say that, for example, a planet in orbit is computing? More specifically, Chalmers (2011) spoke about it in terms of "implementation", asking how we can tell when a physical system is *implementing* a computation. The guiding intuition here was that the state transformations (following the language of Turing's machine) which take place in a rock, or an orbiting planet, do *not* seem to be cognitive in the same way, if at all, as those transformations which take place in the brain of a human, and so do *not* seem to count as computations.

The cognitive processes in question for which functionalism were a means of individuating what counts and what doesn't, were mental "properties" – properties of mental systems – like pain. Functionalism offered a way of individuating mental properties, identifying when they were and were not "implemented" in a system[40], so that philosophers could make sense of questions like "is this AI system experiencing/having/implementing/computing pain?"

[38] For the moment here I am using the terms as of-a-piece.

[39] Authors typically speak of mental "states" in this context. Using the term "processes" is a habit produced my own bias towards an ontology of processes but is any case inconsequential for the point here.

[40]Philosophers like Hilary Putnam who wrote about these questions did not universally use the term "implementation". Its use here is rhetoric, in reference to Chalmers' use thereof in speaking of the implementation of computations.

In order to answer such a question, we need a means of identifying and individuating "pain" or some such mental process. Functionalism is way of doing this[41].

We can now get into the details. Functionalist theories distinguish themselves by individuating a thought or emotion or other mental process in terms of the function of that process in the larger cognitive system of which it is a part. The idea is that the functional role of the process is the causally relevant thing – the difference that makes a difference to whether a mental state is instantiated. By contrast, the substrate or material instantiation of the process is seen as a difference that *does not* make a difference.

The difference between the function and the medium or substrate in which it is realised is both is philosophically and scientifically well recognised. We can think here for example of the difference between the energy realised in the *oscillations* of a wave, or the gravitational contortions of space, and as distinguished from the physical media in which those patterns are expressed. Granted, a wave does not necessarily play a causal role in a larger system, at least not a *cognitive* system, as the notion of cognition is typically understood,[42] so it is important to be careful with the analogies. The point of interest – the interesting thing which functionalism brought to the table - was the decoupling of the material substrate from the causally significant pattern.

In philosophy of mind, again the historically common example has been "pain": the biology of a given human is not strictly necessary as a material substrate for the realisation of pain meaning that, assuming the function of "pain" can be adequately identified, it can in principle be realised in a different material substrate.

The upshot is that, if the function of "pain", or any other process, is physically realised, then the experience will be there, regardless of whether it is realised in a biological or synthetic and artificial system. To be "realised" in this context is typically taken to mean to be causally effective[43], so the functions are causally effective when they are realised physically. As with the wave in water, wind, or space, the same function can be realised in different substrates. The terminology is that functions are "multiply realisable". In technical parlance, it is said that the *causal* effect of the function is metaphysically independent of its *constitution*. For the concerns of philosophy of mind, the consequence is that it does not matter whether the mental state, individuated functionally, is realised in, or constituted by, for example, a carbon or silicon-based substrate. Speaking to the point of interest that motivated much of this way of thinking, Wheeler writes:

> In other words, traditional functionalism provides a principled basis for concluding that creatures whose brains happen to be built out of physical stuff different from our own may still be cognizers. It achieves this heady feat because it bequeaths to the mind the chauvinism-busting property of multiple realizability. To explain: if psychological

[41] Famous alternatives to functionalism in Western, analytic philosophy of mind include the broader Behaviourist tradition, "identity theory" (Place 1956, Feigl 1958), and 'anomalous monism" (Davidson 1970).

[42] The Gaia Hypothesis stands as a renowned example of the proposition that the earth as a planet is a living system and, depending on how far one is willing to take this, a cognitive system, in which case the oscillations of a wave could in principle be considered as part of a larger cognitive system (Lovelock 1979/2016).

[43] The further meaning of "causally effective" is a nontrivial matter in philosophy of mind and entangled in debates about the nature and possibility of "mental causation". These discussions involve wrestling with the conclusion that mental states are epiphenomenal, a conclusion to which we are led if we accept even a basic account of physicalism because, on this assumption, the physical causes involved are those which cause the mental phenomenon. See (Kallestrup 2006) for a summary paper of the issue and the work of Jaegwon Kim (2005, 1998, 1984) for the original work on the matter.

> phenomena are constituted by their causal-functional roles, then our terms for mental states, mental processes, and so on pick out equivalence classes of different material substrates, any one of which might in principle realize the type-identified state or process in question. But of course that means that robots, Martians and the Ood and may all join us in having mental states, just so long as the physical stuff out of which they are made is capable of being organized so as to implement the right functional profiles." (Wheeler 2010: 3-4)

There is a little more worth saying here about multiple realisability because it speaks to way in which this philosophy of mind was cognitivist, or, propagated cognitivist ways of thinking about the mind. Metaphysically, functions and their multiple realisability are a feature of the distinction we can make between *causation* and *constitution*. This metaphysical difference is an important source of navigation in analytic discussions of philosophy of mind in which it is a line relative to which various positions differentiate themselves.

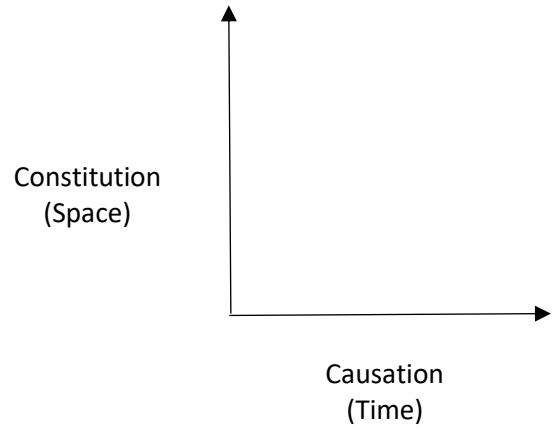


Figure 4. The background metaphysical distinction in terms of which Functionalism is understood.

In the context of cognitivism, it is the fundamental architectural feature of cognitivist metaphysics of mind, where constitution and causation are normally represented as orthogonal axes, with constitution as the vertical (spatial) axis, and causation as the horizontal (temporal) axis.

From a cognitivist and functionalist perspective on these axes, a "function" describes in terms of causation the transformation of an entity or process through time. How that function is physically realised or implemented in space is the matter of constitution. It is this distinction which allows for the possibility of thinking that machines and biological systems are fundamentally the same. Again, this separation is key to the notion of functionalism and will be explored throughout this section.

With that background and context on the philosophy of mind, we can continue to go into more detail on functionalism now.

3.2.3.3. Different types of Functionalism in Philosophy of Mind

As we saw, Putnam is famous for first articulating the position that mental states like pain can be understand as functional states, thereby bringing functionalism to philosophy of mind and the broader developments of the cognitivist model:

> "I shall, in short, argue that pain is not a brain state, in the sense of a physical-chemical state of the brain (or even the whole nervous system), but another *kind* of state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism." (Putnam 1967: 54)

Note the "already-givenness" of the "whole organism".

Putnam expressed his machine state functionalism by analogy to a Turing machine with its well-defined machine states and configurations. The idea was that any creature with a mind can be regarded as a Turing machine whose operations can be fully specified by a set of instructions, as would be written on the "machine table". For Putnam:

> "[a] "machine table" *describes* a machine if the machine has internal states corresponding to the columns of the table, and if it "obeys" the instructions in the following sense: when it is scanning a square on which a symbol s_1 appears and it is in, say, state **B**, that it carries out the "instruction" in the appropriate row and column of the table (in this case, column **B** and row s_1). Any machine that is described by a machine table of the sort just exemplified is a Turing machine." (Putnam 1960: 22) (emphasis in original)

Here Putnam is talking about the relationship between the machine table and its actions. Recall the discussion of the various accounts of computation, which placed different restrictions on when we would say that what is happening here actually counts as computation or execution or an "obeying" of the instructions, or whether the table is just a description.

So, the machine functionalism here is about explaining mental states and their causal relations and action effects by reference to functions that can be specified as they are on a Turing machine table, which is to say, of the form:

> "if the machine in state S_i receives input i_j, then it will go into state S_k and produce output O_l"

In natural language this might translate as, if a mind is in a certain state, receives a certain input, then it will go into another state with some behavioural output, and that is all causally effected. This is the most basic spelling-out of the mind=computer metaphor. Mental states are "machine table states", and the (causal) transformation they undergo is to be described as the description or execution specified on this machine table.

At first this led Putnam to conceive of a mind as a "deterministic automaton", but he later revised this to a "probabilistic automaton" (1967: 54). He evolved his views as the discussion developed. See (Putnam 1988) for his later responses and developments on this earlier work, and (Shagrir 2005) for an overview of Putnam's views.

Some of the challenges of this view have already been discussed in the context on computation. As a reminder, the major problems stem from the indiscriminacy of a 1-1 mapping between machine table states and physical states of the machine. Ned Block and Jerry Fodor raised the "systematicity" and "productivity of thought" challenges in (1972). The productivity of thought challenge raised the point that there are finite possible machine states, by Turing's very own definition, but there are infinite possible states or thoughts for a mind, so a 1-1 mapping like this cannot account for the productivity of thought. This problem of not being able to account for enough of the mind also shows up in the reverse, where it allows for the inclusion of *too much* as a mind – Chalmer's question "does a rock implement every finite state automaton" (1996). As we have seen in the discussion on computation, because a rock could be described with a machine table, does that mean that it is implementing or realising a mind? Taken with the "computational sufficiency" hypothesis, the hypothesis that the implementation of computations (described by a machine table) is sufficient for "mind", one is led to a position in which the rock has a mind. Chalmers articulates an account of "implementation" in (1996) which tries to save Putnam from this problem (one which Putnam himself recognises in the appendix of his *Representation and Reality*). The other challenge, briefly, was the "systematicity" challenge, that there are systematic relations amongst mental states, but that machine functionalism identifies mental states with machine states that have no structural relationship to each other. Recall again that it was this challenge to which Fodor responded by introducing the notion of representation, saying that the structural relations exist between the representations.

Fodor was responsible for articulating the second account of functionalism, psycho-functionalism in his (1968) book *Psychological Explanation*. Where Putnam's machine functionalism was a general theory about mind and cognition, Fodor's was a more scientific functionalism, one that supported the developing science of cognitive psychology. Mental states were individuated and characterised by their role in the theories of cognitive psychology. Of course, mental states are generally individuated and characterised by a scientific theory. In this case, cognitive psychology was the best working theory in the cognitive sciences for explaining human behaviour (Levin 2021).

Compared to Putnam's "is pain a brain state", the question becomes something of the kind "is memory a neural process". Rather than a more general functional state like "pain", more specific cognitive functions like "memory", "attention"[44] and, as will be explored shortly, "vision", are individuated as functions in terms of the role they play for a cognitive system. Note that the existence of this cognitive system into which these functions "bind" as a whole is taken as "already-given".

[44] As a side point, for this system, the neurobiochemistry here is taken to *constitute* or *realise* these cognitive functions. These "higher-order" functions are multiply realisable and the main concern of cognitive psychology and psycho-functionalism rather than "lower-order" biological functions. That these fields chose to focus on this level of the system is nontrivial. Ned Block raised a challenge he called that of "Chauvinism" (Block 1978). The challenge is that of determining the relevant degree of granularity that must be reproduced in order to say that the same mental state is realised. Normally, if a creature shares the appropriate causal patterns which realise the cognitive functions, we say that it is in the same mental state, but what is the relevant degree of granularity? Cognitive psychology and psycho-functionalism focused on the higher-order, coarser-grained phenomena, but it raised the question: if the coarse-grain causal patterns are realised, but not the more detailed ones, do we still say they are in the same mental state?

A third kind of functionalism is called "analytic functionalism". This kind of functionalism is to do with the meaning of our concepts and categories when talking about the relationship between mental states and physical states, the kind of thing Putnam (1960) was responding to. Famous in the philosophy of mind is the question of whether the firing of hypothetical "c-fibres" is identical with "pain", a question first raised by (Smart 1959). Philosophically, the challenge is that "c-fibres" refers to a physical state and "pain" refers to a mental state, so to say that a physical state and mental state are identical is to say that different things are identical, which is analytically false. Pulling on the heritage of Gottlob Frege, David Lewis (1980) responds to this problem by examining a case of a human and a Martian in which they both realise pain in humorously non-identical ways, but because the role or function it plays is the same, we would intuitively say they are both in pain. In this way an analytic functionalism permits that the concepts can have different ways of referring to the same thing.

Putnam's machine state functionalism, Fodor's psycho-functionalism, and an analytic functionalism of the kind just presented, are three major kinds of functionalism in cognitivist philosophy of mind. There is one last pair of functionalisms to present – "role" and "realiser" functionalism – not because they are notably distinct from the three discussed so far, quite the opposite, but because they afford a segue to an important discussion for any account of functionalism, that of Marr and his tri-level hypothesis.

Role functionalism and realiser functionalism make up a distinction discussed in (McLaughlin 2006). Role functionalist theories say that the firing of c-fibres play the functional *role* of pain, whilst realiser theories say that the c-fibres *realise* the pain. The challenge for realiser theories is this. Originally, the power of functionalism was that it permitted us to say that if there are creatures – machines or aliens – with the same functional physical states as us ("c-fibres"), then they have the same mental states as us (pain). It allows us to get past species chauvinism and broaden our philosophising to the "space of possible minds". The problem for realiser theories then is that if there are differences in physical states that realise mental states, something other than "c-fibres", we can't say that other creatures are in "pain". On this basis, most seem to advocate some form of role functionalism.

This distinction between role and realiser, causation and constitution, or computation and implementation, is a functional separation fundamental to the cognitivist paradigm. It is most comprehensively described by Marr in his Tri-Level hypothesis. A brief discussion now will prepare for a discussion of how functionalism the problem-solving conception of cognitivism.

3.2.3.4. From Philosophy of Mind to Cognitive Science: Marr and the Tri-Level Hypothesis

David Marr's "Tri-Level Hypothesis" is first noted in (Marr and Poggio 1976) and then more comprehensively discussed in his now famous 1982 textbook on a computational understanding of *Vision*. See (Marr 2010) for an updated version, (Pylyshyn 1986) for another seminal discussion, and (McClamrock 1991) for a helpful review and critique. Four decades on, the "tri-level hypothesis" does not seem to be discussed all that much in contemporary literature on cognition, in philosophy or cognitive science, with the notable exception of Poggio's (2012) return and revision of the hypothesis. It is something of an "already-given" at this point.

As I will point out, conceiving of and analysing entities in this way is part of what makes cognitivism such a powerful explanatory paradigm, capable of describing entities of all sorts in a coherent, unified framework. The tri-level hypothesis can be thought of, then, as the operative epistemological lens of cognitivism, the fundamental working assumption.

How it matters for my purposes is that this tri-level analysis is functionalist. This is not a controversial claim, but a further distinction is helpful. On one hand there is an "epistemological functionalism", which uses the notion of "function" as a means of cutting up and coming to know the world but which does not necessarily insist that "function" is an ontological kind or real or something like that, and an "ontological functionalism", which would suppose exactly something like that. An ontological sense treats the notion of "function" as a feature of the universe whilst an epistemological sense treats it merely as a feature of humans trying to make sense of the universe.

It is not necessary to imagine that cognitivism or cognitivists believe that function and computation and the concomitant metaphysics are real in the way that an "ontological functionalism" would suggest. Though some, many even, will indeed take the underlying ontology of cognitivism to be Real and True and so on, it is entirely sufficient for present purposes to engage it as an epistemological framework, one which is capable of explaining some things, and less capable of explaining others. So, we do not need to believe that cognitivism or cognitivists think that the tri-level hypothesis is an ontological claim about how the universe is differentiated. So when I say that the hypothesis is functionalist, I am engaging it here as an explanatory paradigm which conceives of entities in terms of the functions they can be said to be realising and executing.

Operating with a general epistemological sense of the term then, recall that functionalism is powerful for talking about how entities solve problems, but, as has been stressed, the identity of that problem is taken as already-given when the cognitive situation itself is taken as already-given. In this way and to this extent, the tri-level hypothesis becomes a very powerful means of talking about how different entities solve different problems, but only because it takes for granted how the problems show up with an agent in the first place. Thusly does cognitivism yield of model of cognition as problem-solving.

The tri-level hypothesis begins by defining the system or process under investigation as an information processing system, and then identifies three separate levels at which the information-processing system can be analysed. Different authors sometimes use different words for the levels. Marr (1982) identifies the levels as *computational*, *algorithmic*, and *implementational*. Computational is the most abstract and implementational is the most concrete, with "algorithm" being somewhere in between. In these terms, the *computational* level is the level at which we analyse what the system is doing in terms of the function and problem it is solving, the *algorithmic* is to do with what rule or process the computations are following or executing, and the *implementational* is the question of how the system is physically realised.

For other examples, Poggio (2012) uses the terms hardware, algorithms, and computations. Pylyshyn (1986) uses the terms semantic (computational), syntactic (algorithmic), and physical (implementational). Sometimes these levels are also more colloquially referred to as content, form, and medium. I have collected the terms in the table below.

Author	Level of Analysis		
Marr (1982)	Computational	Algorithmic	Implementational
Poggio (2012)	Computational	Algorithmic	Hardware
Pylyshyn (1986)	Semantic	Syntactic	Physical
Other	Content	Form	Medium
Philosophy of Mind			Substrate/Realisation

Figure 5. Different names for three levels of analysis of information-processing systems.

Each level identifies a level of functional contribution in the information processing system, whether in terms of "role functionalism" or "realiser functionalism". This kind of thinking produces several important things. Primarily, for the purposes of this thesis, there is nothing in this analysis of the multiple levels of functions which involves defining the information or problem to begin with or accounting for its existence and meaning. The sophistication of the tri-level hypothesis therefore reinforces a sense and sensibility that the relevant way of understanding mind and cognition is in terms of a function. However, this just comes out of thinking with functionalism. One does not need a model like tri-level hypothesis to do that. The model is simply a more sophisticated way of doing functionalism.

An interesting thing that *does* more directly come out of the tri-level hypothesis is a "substrate independence" or "multiple realisability" hypothesis (Putnam 1967: 55)**,** familiar to philosophy of mind and philosophy of AI. It is a metaphysical hypothesis that posits the independence of algorithm and the substrate or hardware it runs on, such that an algorithm or rule can be implemented on more or less any substrate or hardware. It reveals a prioritisation of the abstract pattern or function over the embodied substrate in which we perceive or touch the pattern.

> "Taking the brain-state hypothesis in this way, then, what reasons are there to prefer the functional-state hypothesis over the brain-state hypothesis? Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just a 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." (Putnam 1967: 55-56) (emphasis in original)

Putnam calls out our intuitions that, though something like this might be true, it "is certainly an ambitious hypothesis" (1967: 56).

The prioritisation of the function or pattern over the substrate is not unique to philosophy of mind and AI. It seems to be an idea with heritage in basic physics where, for example, the principle of conservation of energy has us looking "through" a wave to see the pattern and flow of *energy* propagated through the medium of water. Returning for a moment to Shannon and his electrical engineering, Gleick (2012) notes: "Before the engineers quite realized it, they were thinking in terms of the transmission of a *signal*, an abstract entity, quite distinct from the electrical waver in which it was embodied." (2012: 193)

The most general and significant consequence of this prioritisation of function and pattern which comes out of the tri-level hypothesis is the idea that different "functions" or "properties" of human minds, like intelligence or even consciousness, can be realised in different media, including non-biological "hardware" like silicon. It suggests that there is nothing necessarily functionally relevant about the human body with respect to the particularities of the function, pattern, and algorithm(s) of our minds.

This point is nothing less than the metaphysical premise of artificial intelligence research, that the "human is just one means of implementation of the various functions of minds, and that once we can define the relevant algorithms of minds, they can be realised in substrates of more or less any kind." This is stated fairly explicitly in an infamous statement in the project proposal for the "Dartmouth Summer Research Project on Artificial Intelligence" which is often taken as having kicked off the field of AI research proper.

> "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." (McCarthy et al. 1955)

"Simulation" here is the functional realisation of the algorithm or pattern of "learning or any other feature of intelligence" in artificial substrates. Thinking this way allows us to take seriously questions like "does a submarine swim?" which pump a lot of the intuitions that inspire functionalist sensibilities. Insofar as what matters in this light is the function of navigating a body in the medium of water, or something resembling that description, the difference between the organic realisation of "swimming" in animals is no different to the inorganic, technological realisation of "swimming" of submarine craft. They are both doing the same thing. Extending this to "intelligence", in this framework the idiosyncratic means by which organisms "solve problems in a wide variety of environments" (Legg and Hutter 2007) are differences that *don't* make a difference from the exogenous perspective of the scientist – machines realise or implement the same functional capacity. Humans and machines are both solving problems. And now machines are increasingly solving problems to superhuman standards.

3.3. Summary So Far

This marks the end of the discussion of functionalism and, with it, the discussion of the three concepts paradigmatic to cognitivism. I tried to bring out how, in their respective ways, each concept contributed to a problem-solving account of cognition. Together they support a model of cognition which works very well given some important assumptions - what I have been referring to as the "already-givens" of the cognitive situation. It is by virtue of these assumed already-givens that cognitivism can be characterised as focused on problem-solving.

The thing they each and as a whole assume is the cognitive situation itself. The cognitive situation is the situation of an agent situated in a particular environment that it must navigate. When these variables are defined, we have the situation of an individuated agent, situated in a particular context, with a more or less well-defined problem or goal. From this position, the cognitivist model sets to work describing the dynamics of how an agent computes information in order to solve the problem. The agent is composed of an architecture of cognitive functions that work to make this happen. In principle, that architecture need not be human or biological – the "space of possible minds" extends beyond

anthropomorphic examples. Recall the case for the Orthogonality thesis (that more or less any level of intelligence can be combined with more or less any final goal): cognitive functions can in principle be more or less arbitrarily combined.

In the case of computation, I distinguished between abstract computation, as originally expressed by Turing and his machines, and computation in physical systems before summarising a cognitivist conception of computation as "rule-defined information processing" in §3.3.1.3. That computation takes the cognitive situation as already-given is clearest in the case of Turing's machines: the situation begins with an already-given machine processing already-given input information according to already-given rules. The agent is already individuated, the set of information which it processes is already-given in the form of the tape, and the meaning of the input – the particular significance it has for the machine – is also already-given in the form of the algorithm or rule specifying how it is to respond to the input information. Computation therefore seems to describe a process which departs from point where the cognitive situation is already defined, leaving only problem-solving to be done. The question of computation in physical systems makes similar assumptions, albeit with an important difference. Again, there is an already-given situation of an entity in an environment, realising certain transformations (recall the example of the orbiting planet). The central questions are whether these transformations count as computation and what metaphysical conditions need to be satisfied for it to count, and also whether the agent needs to be working with some kind of representations of its environment. However, that basic assumption is still the same, regardless of whether the entity in question is recognised as a bona fide computational agent, namely, the situation of an agent, or entity, in an environment, taking a particular path out of "a choice among possible alternatives" (to segue to a summary of the discussion of information). So, both abstract and physical computation begin as conceptual descriptions of a cognitive process with the cognitive situation already-given, which only leaves for the situation to be played out and the problem solved. It is in this way that computation seems to contribute to cognitivism's problem-solving conception of cognition.

In the case of information, following Aaron Sloman's (2014), I distinguished "control" and "why" information. "Control" information was a reference to Norbert Wiener's use of the term in his 1948 *Cybernetics or Control and Communication in the Animal and the Machine*. Control information is the information of Shannon's (also 1948) information theory, a conception of information as a quantity expressed by a selection or "choice among possible alternatives", with a single "bit" of information being the minimum quantity of a choice between two alternatives (e.g. on/off, fire/don't-fire, heads/tails). I referred to this kind of information as control information because it is the conception of information employed in the design and construction of machines, from Wiener's fire-control anti-aircraft systems to contemporary robotic systems, which are concerned with coordinating their behaviour in feedback loops of control. The key is that this is all independent from any goal the machine may be given, which is to say, the goal of the machine is already-given, as is, again, the machine itself. Information is information only to an already-given entity. Further, when the goal is already-given, it is usually expressed as a rule that affects how the meaning of information is to be interpreted. For example, if the goal is to maximise paperclips, the atoms of humans can take on the meaning (from the perspective of the agent in question) of being raw materials for the production of paperclips. The specification of the goal implicitly defines a way in which to interpret input from the perspective of the agent. By taking the goal as already-given, such a model also takes the meaning of input (again, from the perspective of the agent) as something already-given too. As we will see in the next chapter on post-cognitivism, this is not a necessary assumption to make in order to model cognition, but its effect is that we are left with a view of cognition as goal-independent problem-solving.

Functionalism stands as the third leg of this stool[45]. When employed in a model of cognition, it too takes the cognitive situation for granted. The already-given agent is composed of an assemblage of already-given functions, fit to solve an already-given problem in an already-given environment. Functionalism was articulated initially as a means in philosophy of mind of identifying and individuating cognitive processes in an already-given cognitive agent. The principle of "multiple realisability" holds that such functions can be realised or implemented in different substrates – both machines and humans can realise by different means the same function. Moreover, different cognitive functions can be arbitrarily arranged. For example, intelligence and consciousness need not necessarily be coupled, so the thinking goes. So, anthropocentric conceptions of how cognitive functions are assembled may not be reliable anchors for our intuitions about the kinds of minds that can in principle exist. However, this kind of thinking, in which the existence and arrangement of cognitive functions is taken as arbitrary, is possible if we do not include in our model of cognition how the agent comes to be in the first place. As we will see in the next chapter, the post-cognitivist model of cognition includes the genesis of the agent in a way that shows that the particular composition of the agent is not an arbitrary choice. The point though is that, in making and beginning from these kinds of assumptions, functionalism too is left with an already-given situation of cognition, supporting to the cognitivist view that cognition is just a matter of problem-solving.

Now, for a sceptic, this problem-solving characterisation of cognitivism based on computation, information, and functionalism could, in the last, be dismissed. There are however identifiable symptoms of this problem-solving view and its assumptions. In this final section of this chapter on cognitivism I want to show that both more long-standing foundational theoretical problems in the cognitivist tradition – the frame, binding, and symbol grounding problems – and the more recent questions of the existential risk of AI, can be explained as symptoms of the problem-solving view cognition. Specifically, I will try to show that it is when the cognitive situation is taken as already-given, that the particular challenges of each of these problems arises.

3.4. Fundamental Theoretical Problems of "Problem-Solving" Cognitivism: Frame, Binding, Symbol Grounding

This section is not intended to solve these problems, only to show the way in which they emerge when the cognitive situation is taken as already-given. This is to say, I want to further support my articulation of cognitivism as a problem-solving view of cognition by showing how the fundamental problems of the paradigm are symptoms of taking the cognitive situation as already-given.

Computation and the frame problem are entangled, as are functionalism and the binding problem, and symbol grounding and the notion of information. This may seem obvious, but these three problems are usually taken as separate. With an appreciation of the already-givens of cognitivism that give it its problem-solving character, it is possible to see these three problems as sharing the same root in the ecosystem of thought.

[45] Again, some cognitivists might take the position that representation is a necessary condition for a complete account of a cognitivist view of cognition. However, the problem-solving view of cognition can be made with only these three legs of computation, information, and functionalism. Granted, the fourth leg of representation would offer a more comprehensive "chair", but it does not add anything further to the problem-solving view of cognitivist cognition I am specifically trying to present here.

The frame problem can be shown to be the most fundamental of the three problems – address the frame problem and watch the binding and grounding problems swiftly get in line – so the discussion will begin there. The binding and grounding problems are easier to make sense of with the frame problem framed.

3.4.1. The Frame Problem

The experience of trying to define the frame problem reveals the way in which it applies to itself: what is the relevant way to frame the frame problem? In its broadest interpretation, it is understood as the question of how to specify what is relevant for an agent in a given context (Shanahan 1997, 2016). Defining what is and is-not relevant is not something which explicitly presents itself to us as a feature of cognition, and perhaps this is why it so easily slips through the fences of the cognitivist conception of mind.

Consider the following case. There is a burning building with a child inside and you take it upon yourself to run in and save them. Specifying what is and is not relevant in the context does not present itself as a necessary thing to do and would even jeopardise the safety of the child as we sat there arbitrating (literally a priori) what would and would-not require attention. Armed with courage and sanity, you can be trusted to discern what matters as you go through the situation. It does not need mapping out before the fact.

And the notion of a map is actually a useful one here. A map is a frame on the territory. A useful map shows the relevant features of a landscape - but there are as many possible maps of a landscape as there are ways and "why's" to be moving in the landscape. If I'm in a car, I want the roads to be features, and presented at scale of appropriate to my journey. If I'm alone, hungry, and on foot, without food or water, I want rivers and sources of water, and roads with quick access to civilisation, and so on. The features by which to navigate are whatever the differences which make a difference for me given the situation. If I am journeying in a car, hiking trails are not relevant "differences" or features to be following, and if I am hiking, then motorways are not the relevant features to be following. In both cases, the features or differences continue to exist, but they obviously take on different meaning and relevance, becoming the *relevant* differences in different frames. The situation is thus the frame of reference in which certain features, of a (conceivable) infinity, show up as relevant, taking on the particular significance that they do for the individual, and from their perspective, in the situation. The key cognitivist assumption that takes place here is, again, that the situation is taken as already-given – the frame itself – which means taking for granted what matters to the agent from their perspective. This is fine but, as we shall see, what is in effect happening is that we take the frame for granted, or already-given, and then in our subsequent modelling of how the agent solves the already-given problem, we cannot explain in a non-arbitrary way why the agent encounters certain features as relevant in that context. The frame is the cognitive situation – taking the frame and cognitive situation as already-given leads to the frame problem.

This description is an expression of the frame problem in its most general and philosophical sense. Murray Shanahan is a philosopher and computer scientist who has written extensively about the frame problem (Shanahan 1997, 2010, 2016b, Shanahan and Baars 2005). He defines a narrower and more technological version, which is the way the problem was originally understood when introduced by (McCarthy and Hayes 1969: 30). The problem arose in the context of "classical AI" which, owing to its logic-based approach to building AI, led it to this specific version of the frame problem. Imagine again the scenario with the child in the burning building, but now that a robot is tasked with rescuing

the child, a robot designed, built, and programmable in this classical logic-based way. The narrow version of the frame problem is the challenge of specifying, by means of mathematical logic, formulae that describe the effects of the robot's actions, without having to add formulae ad infinitum that specify the obvious non-effects of those actions. Dennett illustrates the point with characteristic style in his (1984) iterated examples of a robot tasked with getting its spare battery from a neighbouring room, in which there happens also to be a bomb. The first robot is "R1":

> R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT (Wagon, Room, t) would result in the battery being removed from the room. Straightaway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 knew that the bomb was on the wagon in the room, but didn't realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act. (Dennett 1984: 1)

Recognising that the robot must be able to infer not only intended effects, but implied side-effects too, the roboticists build a "robot-deducer" – "R1D1":

> They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT (Wagon, Room, t) it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the colour of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon—when the bomb exploded. (ibid)

The roboticists then realise that the robot needs to be able to distinguish between relevant and irrelevant implications, like the colour of the room's walls; the differences that make a difference for the robot's goals which, of course, are really the roboticists' goals.

> So they developed a method of tagging implications as either relevant or irrelevant to the project at hand, and installed the method in their next model, the robot-relevant-deducer, or R2D1 for short. When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting, Hamlet-like, outside the room containing the ticking bomb, the native hue of its resolution sicklied o'er with the pale cast of thought, as Shakespeare (and more recently Fodor) has aptly put it. 'Do something!' they yelled at it. 'I am,' it retorted. 'I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and...' the bomb went off.

Dennett says he is concerned with the more general version of the frame problem, rather than the narrower version, but his example offers further purchase on the narrower one too. Imagine having to

specify by logical propositions all the relevant and irrelevant differences made by the robot's actions. That is the narrow frame problem.

Shanahan (2010, 2016b) discusses a solution for this narrower version, saying that the "non-effects" of actions can be formally captured if we assume what he calls the "common sense law of inertia" – an assumption that things tend to stay as they are unless acted upon. Formalising this law permits a robot to assume that a feature of the situation, the colour of the walls, remains unchanged by its actions unless it has evidence to suggest otherwise.

In Dennett's last passage above on R2D1, he quotes Shakespeare when he says of R2D1 "the native hue of its resolution sicklied o'er with the pale cast of thought", and then mentions Fodor who also makes reference to Shakespeare in his discussion of the frame problem as "Hamlet's problem viewed from an engineer's perspective", [that is] the problem of "when to stop thinking" (1987: 140).

There is one last remark to make about the frame problem, to do with the ways in which we make inferences. We could say that the frame problem is the challenge of discerning the relevant inference to make, given something from which to make one. What is "informative" about this way of putting it is the suggestion that inferences are "choices amongst possible alternatives". It is not always clear that our inferences, particularly those of the deductive kind, are choices in any nontrivial sense. If all men are mortal, and Socrates is a man, the inference that "Socrates is mortal" doesn't seem like a choice. It seems like a necessity, along with other analytic truths. Participation in the game of logical deduction seems to leave that as the only legitimate possible inference. Not all of our thinking is deductive though.

In the case of Dennett's robots, the robots were making legitimate inferences or interpretations, just not the ones that (external) humans view as *relevant*. Artificial superintelligence is thought to be an existential risk because we can't be sure that, once tasked, it will make the inferences that we deem relevant.

"Aligning" AI with humans is in this light a matter of figuring out how to make what is relevant to humans, relevant to AI. This is something explored in (Cannon 2022) from the perspective of both cognitivism and post-cognitivism.

Failures of alignment happen when there is not an alignment of relevance between the agent in question and the observer, creator or builder thereof. King Midas was not thinking about all the different interpretations and possible inferential paths from his wish. For Midas, the wish-granting satyr did not make, either, the relevant interpretation of his wish or the relevant inference. Again, in the case of superintelligence, if it is tasked with "solve climate change" and then decides to eradicate humans – "no humans, no problem!" – it made an inference that is logically permissible, just not the relevant one from our perspective as concerned humans. It may thereby solve the problem, just not in the relevant way, as we deem it.

Notice two things. Firstly, the way that what is relevant is exogenously specified and defined by the human, a repeating theme of this chapter and the cognitivist model. Secondly, by saying that the inference is not the relevant one, we don't then need to say that it is incorrect or illogical. Anyway, the point is, an appreciation for relevance is thus fundamental to both the narrow and broader versions of the Frame problem. It can now be explained how this is connected to cognitivism's focus on problem-solving.

The problem is the frame, and the frame is the problem. Not in the sense that they are co-extensive though. In a sense that will be explored in the next chapter, a problem can be understood, in effect, as a particular view on the world, one in which features take on particular meanings and significance

relative to that view. More precisely, we can say that a problem, from the perspective of an agent, is their particular frame on the world in which certain things, which might be seen otherwise, take on the significance and relevance – meaning – that they do for that agent.

The frame is a situation from the perspective of an agent, and thus is the condition in which things take on the significance and meaning that they do for that agent. In other words, the frame is the condition in which problems show up *as such*. In each of the sections on computation, information, and functionalism, it was noted that in their respective ways these concepts take the cognitive situation – the frame - as already-given such that all that remains is to solve the problem. Making that assumption amounts to the modeller taking for granted what is relevant or meaningful to the modelled agent. The third and 1st person perspectives are not identical. It is precisely in making this move and taking this for granted that we seem to get the frame problem.

With a certain irony, the frame problem can be understood as the product of a conception of mind which excluded the framing and defining of a problem from its "frame" of what is required to solve that problem. It is an externality of its own making for cognitivism, generated in the model of cognition by what is excluded as part of the cognitive process. We will see in the next chapter that the post-cognitivist theories include in their model how the cognitive situation or frame is generated such that, in principle, the frame problem might not arise in the first place. This is worth further exploration outside this thesis.

Within the framing now established, the binding and symbol-grounding problems can be shown to come from the same problem, at root, as the frame problem.

3.4.2. Binding and Symbol-Grounding Problems are of a Kind With the Frame Problem

The binding problem in particular can be understood as an expression of functionalism, or even as an externality of it. This is clear when we recognise that it too takes the cognitive situation as already-given. Or, more precisely, where the Frame Problem can be seen as a consequence of taking the cognitive situation as a whole as already given, in particular, the binding problem can be seen as a consequence of taking the individuation of the agent in that situation as already-given. The basic format of the problematic is the same though, or, of the same provenance: the binding problem seems to emerge from the basic assumption of the cognitivist model of cognition, namely, the already-givenness of the cognitive situation.

According to this view, the binding problem stems in particular from the functionalism of cognitivism in the following way. In taking the cognitive situation as already-given, cognitivism takes the individuation of the agent in question for granted; this cognitive agent is individuated as an assemblage of cognitive functions (vision, hearing, planning, memory…) and, *because that individuation of the cognitive agent is taken for granted, there is no explanation in the model of the existence of that problem-solving agent and why the particular assemblage of cognitive functions, (those which constitute the agent), exist as a whole.* The binding problem is a thus a problem for the (cognitivist) *model* – "how is it that the features of this agent are bound together into a seamless, unified whole?" Just as with the frame problem in which, because the frame itself – the cognitive situation – is taken in the model as already-given, the model is then challenged with explaining why anything in that situation takes on the relevance that it does, in the case of the binding problem, the individuation of the agent in a cognitive situation is taken as already-given, and so the model is challenged with explaining why the agent exists as it does.

This way of thinking about the binding problem works with a certain interpretation thereof. To substantiate this reading and support my claim that it is a consequence of the cognitivist model's assumptions, I will now go into more detail on how the binding problem is understood in the literature.

At its most general, it is the question of how distinct and separate features of a whole are combined to form that whole. To get into more detail, some basic distinctions are worth mentioning. There are up to four different kinds of binding problems (Roskies 1999, Feldman 2013). For this reason, Roskies (1999) refers to it as a "class" of problems. Like the frame problem, it originated as a much narrower problem, in this case in neuroscience (von der Malsburg 1981). The first problem Feldman (2013) identifies is a general coordination problem: how, for example, are the hands of a violinist coordinated and "bound" together when they are doing two different things. The second problem he identifies concerns our "general unity of perception", that, despite different visual features being processed by distinct neural circuits, our experience is nonetheless unified and whole. The third is the "visual feature-binding" – how the distinct features we perceive in, say, a short, red-haired, beard-sporting gnome, are seen as an integrated whole such that we don't confuse objects which mix those features – e.g. we don't confuse a red circle and blue square with blue circle and red square, or, a blue-haired, clean-shave gnome. The final one is a more analytic problem. The "variable binding" according to Feldman:

> "All animals need feature binding, but variable binding mainly arises in language and other symbolic thought. As a simple case, consider the sentence "He gave it to her before". Four of the six words are variables and need to be bound to values for the sentence to be understood." (Feldman 2013: 6)

This taxonomy is useful for the more neurological of the cognitive sciences. In its philosophical guise, the binding problem is sometimes taken as a subset of the more general mind-body problem – how the body (i.e. central nervous system) responds differentially to different features, but the mind experiences a unified whole. In this way, the binding problem is not accorded particularly significant status in the philosophy of mind. There is a more general philosophical reading of the problem that situates it as an expression of particularly cognitivist models of mind.

Note above in the initial definition of the binding problem it was defined as the question of how *distinct* and *separate* features are combined to form a whole. Both the pre-defined existence of the parts – the features – and the whole are taken as already-given. When their individuation is already-given, there is no way to explain why they show up as an integrated whole to begin with. It is like perceiving branches, twigs, leaves, fruit, and trunk, and then trying to explain why they show up as a whole, when the existence of "tree" is something that was already assumed. Or, more closely to the "modularity of mind" (Fodor 1983), it is the challenge of trying to explain why certain modules are bound together into an integrated whole. In a passage in his *Critique of Judgement*, a passage on which much time will be spent in the next chapter, Kant talks about a watch, in terms of its parts and its whole. He says that the watch does not exist as a whole of itself, but is held together as a whole "outside the watch" in the mind of one for whom the whole can be a causally effective thing:

> One part is certainly present for the sake of another, but it does not owe its presence to the agency of that other. For this reason, also, the producing cause of the watch and its form

> is not contained in the nature of this material, but lies outside the watch in a being that can act according to ideas of a whole which its causality makes possible. (Kant 2007: 202)

This brings up again the importance of noting the 3rd person perspective on an object that it implicitly taken by cognitivists going back to Shannon – it involves an external, exogenous agent who defines the watch as whole of bound parts, and external, exogenous agent *who is not included in the model of cognition*. The same situation happens in the case of individuating cognitive modules as a whole "mind", neural circuits as a "brain", and perceptual features and variables as an "object". When the cognitivist model takes the watch as a whole as already-given and decomposes it into separable functions, there is no way to get from the functions back to the whole. There is no non-arbitrary way of putting the functions together to make the whole. In its positive guise, this point is the source of the Orthogonality thesis: a goal is functionally orthogonal to a cognitive capacity to achieve it. In its negative guise, this point looks to be the source of the binding problem: each specifiable function can be treated as independent of one another. In the language of Kant's quote above, they "owe" nothing to each other, and so we are left with the question of how they seamlessly bind together.

By contrast, in the post-cognitivist model of enactivism I present in chapter five, the story of cognition is a story of how the agent comes to be, and be in the situation that they are. The individuation is not arbitrary in the model and so, no issue of binding arises in the first place for the scientist.

The binding problem is therefore of a kind with the frame problem for the way in which it is confronted with explaining what the (cognitivist) model takes as already-given. Both are problems to do with already-givens which cognitivism assumes for its picture of cognition.

The last of the three problems foundational to cognitivism is the symbol-grounding problem. In its original exposition, Stevan Harnad (1990) defines the symbol-grounding problem with the question: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?" Harnad's formulation expresses an awareness of the distinction between 3rd and 1st personal perspectives and advocates for a 1st person perspective, the perspective "intrinsic to the system." In order to explain the grounding problem, Harnad leverages Searle's (1980) "Chinese room" thought experiment and Turing's "imitation game" (Harnad 1990: §2). In both of these cases he (Harnad) notes what was understood at the time, that symbol manipulation is insufficient for meaning. Exchanging symbols for other symbols may suggest functioning, but not comprehension. He suggests that this is what the grounding problem is. He further illustrates it with an invitation to imagine learning Chinese as a second language by means only of a Chinese/Chinese dictionary. "The trip though the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol or symbol-string to another, never coming to a halt on what anything meant." (Harnad 1990). The use of "halt" is worth double-clicking on. The paradoxes mentioned in the context of the earlier discussion of Turing and the Decision problem and Halting problems were cases of "never coming to a halt on what anything meant" because if it meant one thing, then the contradictory thing was proved true and so on. (e.g. "This sentence is false.")

Harnad then takes his thought experiment one step further:

> Suppose you had to learn Chinese as a first language and the only source of information you had was a Chinese/Chinese dictionary! This is more like the actual task faced by a purely symbolic model of the mind: How can you ever get off the symbol/symbol merry-

> go-round? How is symbol meaning to be grounded in something other than just more meaningless symbols? This is the symbol grounding problem. (Harnad 1990: §2.2)

The symbol-grounding problem originally concerned symbolic, "classical AI", and whilst there are now many ways to deal with it, (see Taddeo and Floridi 2005), from the perspective of this chapter, the grounding problem arguably concerns any computational system exchanging or "communicating" signals and information. For any such system we can ask, how do those signals, whether symbolic, or non-symbolic (electrical potentials in brain, train signals, or any non-symbolic "control" information) become meaningful for the agent from their perspective? That is, how do they come to take on and effect the particular significance that they do for the agent?

I have described these problems as the fundamental theoretical problems of cognitivism. This is a relatively strict characterisation and excludes other well-known problems in philosophy of mind – Searle's "Chinese Room" (1980) for example – which might legitimately be included in the category of fundamental theoretical problems of cognitivism. I think problems such as the "Chinese Room" and related problems can be shown to be symptoms of the same cognitivist set up. Showing this would be valuable work, but would overextend this thesis. The point was to show how the frame, binding, and grounding problems emerge from the basic assumption of the cognitivist model of cognition, the already-givenness of the cognitive situation. These problems are representative of the range of disciplines in cognitivism, including the cognitive sciences and philosophy. In this regard they are hopefully sufficient then.

I want to finish this section with a small commentary. Seen in the light of cognitivism's already-givens, the frame, binding, and grounding problems stand in interesting relation to one another. *We can say that the frame problem is the consequence of taking the cognitive situation as already-given. We can say that the binding problem is the consequence of taking the whole or agent in question as already-given. And then the symbol-grounding problem can be seen as a consequence of taking both as already-given, that is, as taking an agent in a context as already-given.*

More generally of these three problems, it seems that as long as the mode of epistemological inquiry is conducted from the 3rd person perspective, modelling meaning for the agent from the perspective of that agent will always be a challenge. It amounts to trying to model from a 3rd person perspective a by-definition 1st person thing.

This section on the "fundamental theoretical problems of cognitivism" was intended to draw a connection with the already-givens shown in the discussion of computation, information, and functionalism. Collectively the already-givens were just the cognitive situation – a problem (frame), a problem-solver (information-processor), and the relevant information for that agent to solve the problem. *When these things are already-given, they are excluded from the explanatory power of the model and show up as externalities – problems for that model – later on.*

It is worth considering that, in this light, one could equally view these externalities as *discoveries* of cognitivism. In a plot twist we could say "Cognitivism has explored and discovered, by learning of the externalities, the importance of taking the cognitive situation as already-given."

3.5. Conclusion

In this chapter, I aimed to present the fundamentals of the cognitivist paradigm of mind in order to support my characterisation of cognitivism as a "problem-solving" view of cognition.

I began first by noting that "cognitivism" is not actually a term by which this paradigm defines itself. Authors who share the fundamental assumptions of this paradigm do not specify they are cognitivist each time they write a paper. Instead, "cognitivism" is a name given mostly by those (e.g. Varela et al. 1991, Wheeler 2005, 2008, 2010, Thompson 2007) who do not subscribe to the same fundamentals. Where authors have distinguished "cognitivism" as the earliest stage of the paradigm and, for example, differentiated it from connectionism or embodiment approaches to cognition, I have included all of these under the heading of cognitivism because they share the same fundamental assumption in their respective models of cognition. I have laboured to suggest that this assumption is no less than taking as already-given the cognitive situation in question.

In order to show this, I identified the main concepts of the cognitivist model – computation, information, and functionalism – and pointed out that the work they do for the cognitivist model variously requires making the same assumption about the cognitive situation. I excluded the notion of "representation", which many would characterise as distinctly cognitivist. I made this exclusion on the basis that, primarily, representation is unnecessary to substantiate the problem-solving characterisation of cognitivism that I am making and, secondly, that representation emerged anyway in the work of Fodor (1975, 1981) *after* computationalism and, arguably, in order to support a computationalist conception of cognition that was already in place[46]. This is to say, representationalism emerged after the initial assumption of cognitivism had already been made.

I began the chapter then with computation first. I discussed the abstract sense of computation first, presenting the notion of Turing Machines as the most basic articulation of the computational aspects of the cognitivist model of mind. In order to show how computation contributes to a problem-solving perspective, I emphasised that the notion of computation assumes the already-givenness of the cognitive situation: there has to be an agent or entity in a world with a problem in order for there to be something computing. I suggested that the conceptions of computation in physical systems distinguish themselves for being more precise about what counts as computation proper as opposed to just change or transformation of some kind, all such that we can say that all and only the things compute which we actually want to say compute, i.e. not rocks. The notion of computation in physical systems otherwise shares the same set up and takes the cognitive situation as already-given.

I then moved into a discussion of information. Beginning with Gregory Bateson's "the difference that makes a difference", and then focusing on Shannon's quantitative theory of information, I used Sloman's conception of "control" information to describe how information is understood in cognitivism. This particular conception proved more valuable in the context of this chapter compared to a more prevalent "semantic" – "syntactic" distinction. The distinction between syntactic and semantic does not explain why machines struggle with semantics, but recognising that the kind of electrical engineering systems with which Shannon was concerned were basic control systems (from basic circuits to anti-aircraft weapons systems), reveals that "control" information is all that is necessary for

[46] This is not something for which I made a detailed argument. It is a place in this thesis where I am making an important assumption. That said, the legitimacy of the problem-solving characterisation of cognitivism I have made in this thesis does not turn on the inclusion or exclusion of representation. I aim to have shown in this chapter that the problem-solving characterisation is a legitimate and generative way of understanding cognitivism and that, to do this, computation, information, and functionalism are sufficient.

coordinated, goal-oriented behaviour. I pointed out that "control" information is independent of what we could be called "why" information", the kind of information which involves meaning for a goal or action might be worth taking in the first place. In this way, the operative conception of information contributes to the problem-solving orientation in the cognitivist model because it too assumes the problem as given and only contributes to an explanation of how the problem is solved. Information as a concept does not account or contribute to an account in the cognitivist model for how the agent's problem comes to be meaningful or significant to the them from their perspective.

The last paradigmatic concept was functionalism. I began with a discussion of functionalism in philosophy of mind – Putnam's machine functionalism, Fodor's psychofunctionalism, and then "analytic functionalism". This was supplemented with a discussion of functionalism in cognitive science in which I considered Marr's Tri-Level Hypothesis. I made the point that functionalism has a unique strength (that it is a framework which enable us to talk more or less universally about any entities in terms of information processing) and weakness (that it struggles to speak to why particular things are relevant for particular entities). This is the case, I claimed, because of the way the specification of a function (in a model) excludes the genesis of the function in that model. The function explains (in a model) *how* a certain process happens, but *why* the execution of the function matters, to the agent in question, or to the modeller, is taken as already-given. A function is a useful tool for modelling how a problem is solved, but has to taken as already-given why that problem is worth solving, to the agent in question, or to the modeller. Functionalism is therefore distinctly appropriate for, and limited to, a problem-solving view of cognition.

I wanted then to go further. I wanted to show that not only can we see cognitivism's problem-solving character in its fundamental concepts, but that it is also apparent in the fundamental problems it has as a paradigm. This led to a discussion of the frame, binding, and symbol-grounding problems. To repeat the conclusion, the frame problem is the consequence of taking the context – frame – as already given, the binding problem is the consequence of taking the whole or agent in question as already-given, and the symbol-grounding problem can be seen as a consequence of taking both as already-given, that is, as taking *an agent in a context as already-given.*

Before turning to post-cognitivism and seeing what that model looks like, I want to take one last step in this discussion of cognitivism. In this thesis I want to show two things. I want to show that cognitivism and post-cognitivism have two different ways of thinking about cognition and therefore, whether AI can become more ethical than humans. However, I also have the broader concern in mind that our existing thinking about AI is distinctly cognitivist. In order to show this, I now want to show how the same cognitivist assumption I that shows up throughout the cognitivist paradigm, also shows up in current thinking about AI, the question of superintelligence and existential risk in particular.

4. The Problem-Solving Cognitivism in AI Existential Risk: Possible Minds, Superintelligence, and Orthogonality

The broadest aim of this thesis is to show that when it comes to thinking about artificial intelligence, there are in fact two available paradigms of mind research on offer, and that existing work on AI is uniquely and exclusively rooted in the cognitivist model. In the first chapter I situated the thesis in the context of the literature on the claim that AI represents an existential risk to humanity. In the following chapter, where we have come to so far, I presented the foundations and core tenets of the cognitivist model of mind, repeating the claim that the way the model takes the cognitive situation as already-given yields a problem-solving conception of cognitivism. I now want to put these things all together to show how the claim that AI is an existential risk is, again, uniquely and exclusively a product of the cognitivist model.

I begin here with a reminder of how AI is thought to be an existential risk (xrisk) and then show how a key support for the AI xrisk claim, the "space of possible minds", is a construct proper to the cognitivist model of cognition. I then focus my sights on indicating how the xrisk claim is a product of computation, information, functionalism, and the model they jointly amount to. Before turning at last to the post-cognitivist model and by way of summarising the content of thesis thus far, I finish with a short dialogue between a proponent of the (cognitivist) view that AI is an xrisk and someone representing everyday intuitions about intelligence and minds.

4.1. AI Existential Risk Recap

Quickly running through the AI xrisk concepts and narrative again, existential risk (xrisk) is a category of risk pertaining to the existence of humanity (and much else). To say that AI is an xrisk is to say therefore that AI is a risk to the existence of humanity.

The story of how we go from existing systems to systems which represent this kind of risk does in fact begin with a regular AI system. As this system goes about solving the arbitrary problem with which it has been tasked, it realises among several other things that if it were more intelligent, then it could better solve the given problem, and so divests resources to self-improvement. Recursive feedback loops of improving self-improvement mean that at some point the once innocuous AI system rockets past "human-level" intelligence into superintelligence. What happens at this point is beyond the horizon of perception for us, hence the name "Singularity". The risk to humans comes from the idea that the now superintelligent AI will still be pursuing the original, humble goal with which it was purposed, hence the "singularity paradox" – because one might think, given greater intelligence that superintelligence might have greater aspirations. However, these two things are understood to be independent and "orthogonal" such that superintelligence can still pursue a trivial goal. What makes it a problem and risk for humans, despite the triviality of the goal, is that the superintelligence now is capable of marshalling resources and conducting affairs in a way, and at a scale, which puts it at odds with humans. By analogy, it resembles the way building motorways is at odds with the ecosystems over which they pave. In whatever way humans and superintelligence might come into rivalry, even if only in competition for instrumental resources, superintelligence being by definition orders of magnitude more intelligent than humans means that humans are powerless to stop it. Moreover, the demands of instrumental intelligence mean that a superintelligence would, in efforts to ensure the optimisation of

its goal, figure out ways to protect itself from being shut down, ways that could conceivably include sophisticated modelling and anticipating of human behaviour such that we cannot just "pull the plug".

The particular and given goal is not necessarily the problem. It is the mere fact of superintelligence being orders of magnitude more intelligent and potentially coming into a rivalrous dynamic with humans. For almost any "final goal" a system has, there is a set of goals which are instrumentally necessary, so-called "instrumental goals". It is via these instrumental goals - things like "self-preservation", "resource-acquisition", "self-improvement", and so on - that ASI becomes existentially dangerous. Humans and superintelligence share instrumental goals which means that we may come into competition with superintelligence. It is in this potential rivalry with ASI that xrisk to humans arises[47].

The claim of this chapter is that the narrative that AI represents an existential risk is based on the cognitivist model of cognition. Before engaging the narrative, I want first to tackle the shield it wields as defence. Recall from section 2.1.3.2 on the "Space of Possible Minds" that authors like Bostrom and others defend the plausibility of exotic minds from intuitions of the kind that "that no creature that intelligent would seriously devote its cognitive powers to any arbitrary goal" by saying that in the space of possible minds, minds like human minds occupy only a small corner of what is possible; further, they argue that we should not anchor our ideas about what is and is not possible in anthropomorphic intuitions about minds. The space of possible minds is supposed to, in this regard, be a basis for avoiding being misled in thinking about minds. In the first section of this chapter however, I want to show that it is inextricably rooted in functionalism about mind. In this regard, it has not identified the constraints or architecture of *any* possible minds, but minds of a distinctly cognitivist kind. The point is to show that, *as a basis for defining what kinds of minds are possible*, the notion of "the space of possible minds" is distinctly functionalist, and therefore inextricably cognitivist.

With that in place, I will then point out how the notion of the existential risk of AI falls out of the key notions of cognitivism I discussed in the previous chapter: computation, information, and functionalism.

4.2. The Space of Possible Minds as a "Frameless" Functionalism

As a reminder, the "space of possible minds" was first articulated by Aaron Sloman (1984) and then later picked up and employed in the work of José Hernández-Orallo's *The Measure of all Minds* (2017). The idea is that in the relative infinitude of the space of possible minds, "human", and even "human-like", minds occupy a tiny niche of the possibility space, so it may not be right to expect that the particular "framings" and "bindings" of functions particular to humans – e.g. consciousness + intelligence –generalise or quantify over much of the possibility space. Put more simply, human minds are not the only conceivable kinds of minds. Moreover, when we consider all the possible combinations of even just the functions we attribute to human minds (consciousness, intelligence and whatever else), it is clear that the particular human combination and binding is a small subset of the space of possible combinations and bindings, presumably particular to the evolutionary and ecological constraints and selection pressures by which the human mind has been moulded. Who knows what other alien and artificial functions could be attributed to something we might consent to call "mind". Allowing for this possibility, the space becomes even more exotic.

[47] Return to chapter 2 for a refresher. See also (Müller and Cannon 2022) for an analytic reproduction and critique of the argument that AI represents an existential risk.

I originally brought it up it in the second chapter because is often cited in the context of AI xrisk to remind us, when speculating about the possibilities of AI, that we ought not anchor ourselves in anthropomorphic models and conceptions of mind (Shanahan 2016a). Whilst this is an insightful point, it is nonetheless characteristic of cognitivism. In particular, I want to claim that it supposes a functionalism distinctive of cognitivism: the space of possible minds might well be called "the space of possible functions".

To understand this, consider again one aspect of the binding problem discussed at the end of the previous chapter – why is it that certain different, separable, cognitive functions show up unified in humans, (and how does that happen)? The space of possible minds is the idea that all these functions can in principle be combined in more or less any way. Recall that in the cognitivist model of cognition, functionalism serves to explain *how* a problem is solved, but not *why* that problem is a meaningful problem to solve (from the perspective of the agent) in the first place. The claim in the previous chapter was that this distinction arises when the model in which the functions are described take cognitive situation, (and the agent in particular), as already-given. I spoke about Kant's watch, in which he points out that the "whole" of the watch lies outside the watch in the mind of the maker – the modeller in this case.

The space of possible minds permits discussion of some counterintuitive ideas, and functionalism shows up in each case. In the paper which inspired the question of this thesis, Bostrom and Yudkowsky (2008) go so far as to speculate about the "decoupling" of "sentience" and "sapience" (consciousness and intelligence, respectively), treating each as a modular function that, without an endogenous notion of a whole, are only together for an arbitrary reason. This is to say, the agent constituted by these functions is taken as already-given.

Continuing this line of thought, notice that thinking about the mind in terms of properties like this eventually lends itself to the idea that some of the properties are independent or "orthogonal" to one another, and that, in the space of possible minds, there is no a priori reason why certain properties necessarily have to be coupled. There is a lot of mileage to this kind of thinking, particularly in the philosophy and ethics of AI. Exploring the moral status of hypothetical machines with exotic combinations of properties stimulates a lot of theoretical research, including this thesis. In this way, the binding problem is treated not as a "bug" but is actually an expression of a "feature" of this way of thinking – it leads us to imagine an agent with independently variable values of, for example, intelligence and consciousness.

The point is that such philosophical speculation only makes sense given a basic functionalism in which we understand the human mind to be a contingent coupling of functions which help humans solve an idiosyncratic and ontologically-arbitrary problem-set, particular to our evolutionary phylogeny and/or ontogeny. If the assemblage of these particular functions naturally seems arbitrary, then it is legitimate to suppose different arrangements and assemblages of those problem-solving functions we call mind and cognition. However, it is arbitrary only when it is taken as already-given.

Put another way, given an understanding of mind and cognition which sees it as the execution of functions bound together, then it is makes sense to explore what happens if we imagine separating those functions and putting them back together again in different combinations. Below is a graph from Shanahan's *Aeon* essay *Conscious Exotica* (2016a) in which this kind of thinking is clearly expressed.

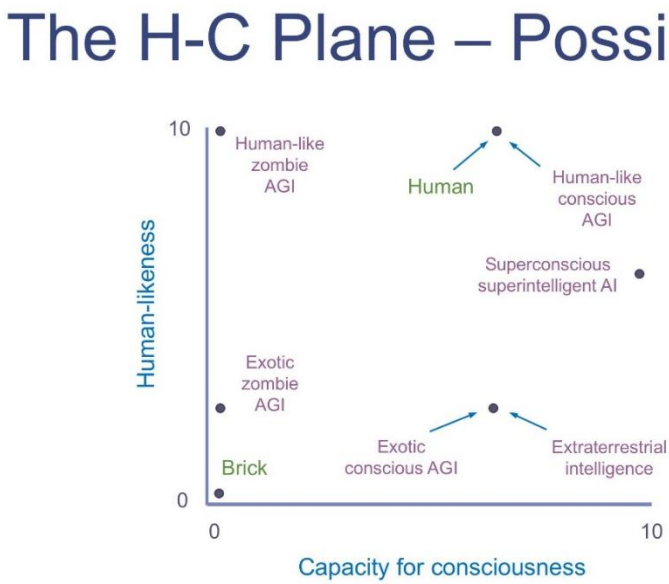


Figure 6. Shanahan's (2016a) exploration of different possible "bindings"

Weird and exotic entities with exotic competences emerge with such re-arrangements and re-assemblages, things which seem paradoxical to human minds (Shanahan 2016a). Note again that "entity" is taken for granted in the same way that "AI" is. The "whole" as a referent is taken as already-given. Here the whole is individuated in terms of a linear, sum-of the-parts, part-whole assemblage of functions or properties.

The "Singularity Paradox" mentioned in the second chapter, in which an artificial superintelligence is both superintelligent but also seems super stupid, can also been seen as an expression of the arbitrary and exotic "binding" of functions. Here the arrangement of cognitive functions is such that artificial superintelligence is simultaneously superintelligent and super stupid insofar as it will devote its super problem-solving capacity to a super banal goal/problem – nothing in the notion of intelligence, mind, or cognition involves defining or choosing the problem. These things are already-given.

The way in which such paradoxical conceptions of mind emerge from a notion like the space of possible minds is interesting, particularly following the discussion of the frame problem in the previous chapter. The space of possible minds can be understood as the consequence of removing the human as a "frame" of reference through which possible minds can be conceived. The space of possible minds is in this way "unframed". The reason in favour of removing the human perspective in the first place, namely, to avoid biases of anthropomorphism, makes sense, but it then also makes sense that a model of frameless minds should face the Frame Problem.

The heritage of this thinking goes back to cybernetics. In his excellent chronicling of the birth of cybernetics through the ten Macy Conferences held between 1946 to 1653, Jean-Pierre Dupuy (2000), points out that the first cybernetic movement, which gave birth to cognitivism, "provided the formal means for conceiving the category of *subjectless process*" (ibid: 156), or, more specifically, "subjectless

cognition"[48]. There is no subject from the perspective of which the cognition takes place. It is, as such, without a frame.

Frameless, subjectless, de-anthropomorphised cognition, was a path developed and taken to make advances towards a more ontologically universal, rather than "local", perspective on cognition. But in removing any frame, cognitivism came to the Frame Problem. In short, cognitivism created its own problems. This was a point on which I concluded the discussion of the fundamental theoretical problems of cognitivism in the previous chapter.

It is what happens when, in taking the cognitive situation as already-given, we exclude from our model the context, relevance of the problem that a cognitive agent is trying to solve. The space of possible minds is like the walls and drawers of a garage filled with tools. Like a function, we are thinking with an orthogonality here - "more or less any [tool] could in principle be combined with more or less any final goal". With the space of possible minds, we are engaging in a survey of the decontextualised "tools" in the garage, and imagining the space of their possible combinations. But we saw with the screwdriver that what, in Shannon's terms, defines the set of "possible alternatives" and then the "choice amongst possible alternatives" in this space, *is* the frame or context. This is what is meant by the space of possible minds being "unframed" – there is not frame or context. As such, tools and functions can then "in principle be combined with more or less" any other.

It stands to reason then that it would reveal a landscape of exotic, "artificial", and paradoxical minds. It also stands to reason that, absent a frame of reference, we will struggle with such a model to account for how and why anything is relevant for such exotic and artificial minds from their perspective. This seems to be exactly what is happening where the paradoxes show up. For example, in an unframed domain of possible bundles of independent and orthogonal functions, superintelligence of the "singularity paradox" kind is a logical possibility.

The space of possible minds sits as a foundation stone in the edifice of the structure that is the claim that AI represents an existential risk to humanity. Having now claimed that it is rooted in a "frameless" functionalism distinct of cognitivist assumptions about cognition, I will now turn to the rest of the xrisk story to claim that it falls out of the key cognitivist concepts I have been discussing.

4.3. AI Xrisk Falls out of Notions of Computation, Information, and Functionalism

The most relevant concepts in the existential risk narrative are Superintelligence and the Orthogonality Thesis. These two and the broader xrisk narrative require the cognitivist conceptions of computation, information, and function in order to work. As parts and a whole, these concepts are as much products and conceptual affordances of cognitivist ecosystem of thought as anything else. The key thing is the

[48] It is important to note that Varela et al. (1991: 105-130) speak of "selfless-minds", which gives the impression of being something similar, and Dupuy does identify a connection (2000: 160), but there is a subtle and significant difference. The work of Varela et al. is of the "post-cognitivist" paradigm to be discussed in the next chapter and has important links to contemplative traditions, Buddhism in particular, in which context the "self-less mind" is so *despite* the experience of being a situated, perspectival self. (See also Thompson 2015). The "subjectless cognition" of cybernetics to which Dupuy makes reference doesn't have a "self" to see past, the way contemplative Buddhism might speak of it. In "subjectless cognition", there is never a self to begin with.

same thing that has been repeated, namely, they are based on a model which takes the cognitive situation as already-given.

The way this shows up is that computation, information, and functionalism variously explain and participate in explaining in a model *how* the cognitive process of an agent occurs. However, because the problem or goal is already-given in the model, accounting for the meaning of the problem or goal to the agent in question *is not a feature of the model.* It is excluded. This part of the cognitive process – and we have reason to believe that it *is* a part of the process, based on everyday experience and on the post-cognitivist model – is excluded from the model. It stands to reason, then, that, in reasoning according to this model about how an agent, (like a Superintelligence) might behave (e.g. Omohundro 2007, 2008 and Bostrom 2012, 2014), the part of cognition in which we discern problems and goals is not a feature of how the modelled agent behaves. We see exactly this in the description of Superintelligences which blindly execute banal final goals.

Looking in more detail now, we see it in particular in the way in which intelligence is conceived. Recall that it is conceived in the xrisk narrative as "instrumental intelligence" (§2.1.2.) As a reminder:

> For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning … Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. (Bostrom 2012: 73)

> "By "intelligence", we here mean something like skill at prediction, planning, and *means-end reasoning* in general. This sense of *instrumental cognitive efficaciousness* is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be." (Bostrom 2014: 107) (emphasis mine)

Again, note that this instrumental conception of intelligence treats the final goal as already-given. This conception of intelligence is all but identical with a conception of cognition as computation in several ways. Fundamentally, both are goal-independent: recall from the section on computation (§3.3.1) that computation is function or goal-independent because it can be performed in the service of any function or goal. Both instrumental intelligence and computation can be in service to "more or less any final goal".

This in turn begins to make sense of how a notion like the Orthogonality Thesis emerges. When the goal is already-given in the model, there is no need to account for how the agent comes to discern or acquire it. Moreover, when the goal is well-defined, which it often is not, the capacity to achieve it is then in principle amenable to infinite scaling because well-defined goals can become optimisation problems.

By contrast, it is hard to know what to optimise for if the problem is not clear. For a conception of intelligence as instrumental intelligence, rooted in a conception of computation, both of which must assume that the problem is well-defined, it stands to reason that such notions will produce a conception of intelligence which defines it effectively as an optimisation process:

> "We use the term "mind" here simply as a synonym for an optimizing agent. … In our understanding, any optimization process, including a hypothetical artificially intelligent agent above a certain threshold, would constitute a mind. Nonetheless, the intuitions for the concepts of "mind" and "intelligence" are bound up with many human properties, while our focus is simply on agents that can impact our human future. For our purposes, then, the terms "mind" and "intelligence" may simply be read "optimizing agent" and "optimization power." (Yampolskiy and Fox 2012: 3)

> "The notion of an "optimization process" is predictively useful because it can be easier to understand the target of an optimization process than to understand its step-by-step dynamics." (Yudkowsky 2008: 10)

Once we assume the nature of the problem and define a function for solving it, AI can and does optimise beyond the "human-level". This is to say, instrumentalised intelligence as conceived in terms of computation in principle admits of infinite scaling because, once the problem is defined, we can throw more data and compute power at it such that the same problem is solved faster and faster. This in turn begins to make sense of the notion of Superintelligence.

In the second chapter in which I discussed the xrisk context of this thesis, I explored a navigational metaphor for making sense of intelligence. If we imagine terrain we have to navigate, moving from point A to point B, intelligence in this picture is the navigational problem of getting from A to B. Solving the problem amounts to finding the optimal path through the space. "Optimal" could of course be fastest, cheapest, most efficient, most scenic, or any other variable for which we might want to optimise, "framing" the problem such that what is valuable or maximises "points" is clearly defined. Of course, whilst such a metaphor has a lot of mileage to it, it too requires taking the problem as already-given: "getting from A to B, optimising for X".

In a context where intelligence is understood as this manner of instrumental optimisation, it makes sense to imagine that, with an already-given problem, something like a super-optimiser (aka superintelligence) is possible. At least, it is not clear why there would be any particular limit, other than what is defined by the laws of physics, that determines what level of optimisation is possible. The architecture of the space of possible optimisation functions is not defined by the constraints of anthropic minds. Where computation and instrumental intelligence identify the same thing, superintelligence seems to be an expression of that same thing, just taken to the extreme.

The claim then is that computation and instrumental intelligence therefore amount to the same thing. Not necessarily metaphysically, but at least *functionally*, namely, solving a problem without regard to what the problem is. This falls out of the cognitivist assumption of taking the cognitive situation as already-given. It is an important, implicit detail which gets us from this basic notion of instrumental intelligence to artificial superintelligence.

Even alone, an understanding of "computation" goes a long way to understanding the xrisk narrative, and superintelligence in particular. However, the functionalism of cognitivism is also at work. Consider the Orthogonality Thesis again:

> "Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom 2014: 107)

Note the way that achieving a final goal is decoupled as a cognitive capacity from any kind of discernment of a meaningful final goal. They appear to be *treated as separate cognitive functions*, in line with cognitivism's broader functionalism. Recall a passage from second chapter in which Bostrom uses this kind of language:

> "As previously indicated, we use the term "superintelligence" to refer to intellects that greatly outperform the best current human minds across many very general cognitive domains. This is still quite vague. Different kinds of systems with rather disparate performance attributes could qualify as superintelligences under this definition. To advance the analysis, *it is helpful to disaggregate this simple notion of superintelligence by distinguishing different bundles of intellectual super-capabilities. There are many ways in which such decomposition could be done.* Here we will differentiate between three forms: speed superintelligence, collective superintelligence, and quality superintelligence. (Bostrom 2014: 52) (emphasis mine).

Here Bostrom expresses his thinking with exactly the kind of fundamental functionalism that has been under discussion. Note in particular the proximity to the binding problem. In the space of possible minds, its strength comes from suggesting that the human arrangement or "binding" of cognitive functions need not be the only way to do things, and that other alien and artificial Frames are conceivable which might bundle different functions together. In the above quote, Bostrom does the same thing. He is playing with the possible arrangements of different cognitive functions or "super-capabilities", each admitting of super-optimisation, to explore different forms of superintelligence. Bostrom echoes Fodor's modularity of mind, and faces the same binding problem present to its functionalism.

The Orthogonality Thesis is the idea that, in principle at least, some functions which are bound together in humans are really independent first. In this way, the Orthogonality Thesis makes the same move that leads to the binding problem. In taking the "whole" as already-given, there is no available means of explaining in a non-arbitrary way why the separate functions cohere together. Just because there is an a priori, analytic, conceptual separability of "intelligence" and "goals", sentience and sapience, does not mean that the referents themselves are separate. Conceptual separability does not necessitate "in fact" separability[49].

In the end the moves are fairly straightforward. On the one hand, supported by a model of computation, we have a conception of intelligence as instrumental intelligence, a conception in which the final goal is already-given and well-defined, and any kind of discernment at this level about what goals are *worth* pursuing is excluded. On the other hand, we stipulate a thesis which claims that these two things are orthogonal. Put together, it is a short step to imagine that instrumental intelligence applied to an already-given goal could be scaled arbitrarily and so we get "Superintelligence".

[49]As Dennett puts it: "One should be leery of these possibilities in principle. It is also possible in principle to build a stainless-steel ladder to the moon, and to write out, in alphabetical order, all intelligible English conversations consisting of less than a thousand words. But none of these are remotely possible in fact and sometimes an impossibility in fact is theoretically more interesting possibility in principle..." (Dennett 1991: 4)

The way in which AI xrisk falls out of cognitivist conceptions becomes quite clear then[50]:

1. Stipulate the independence of a capacity to achieve a goal from a capacity to have or discern goals in the first place. (Orthogonality)
2. Stipulate that intelligence is only one of these, the capacity to achieve a goal. (Instrumental Intelligence).
3. Conclude that this instrumental intelligence admits of in-principle infinite scaling to superintelligence. (Superintelligence)

There is a conspicuous necessity to the way these ideas fit together, producing the proposition of a Superintelligence that does not have a capacity to reflect on its final goal and so instrumentally terminates humanity in order to achieve whatever arbitrary goal it has. It is conspicuous because it is unerringly logical given the way the terms are defined, but, again, that is something that is not brought into question. This line of thinking doesn't happen without restricting a notion of intelligence in this way, a restriction to "problem-solving".

A quick objection to the above 1-2-3 should be noted. It might be said that Superintelligence is not simply a scaling of instrumental intelligence, and that the kind of Superintelligence that is an existential risk is a domain-general "general intelligence", as opposed to domain-specific and "narrow intelligence" (of the kind of a calculator). If this is the case, something else is necessary between 2 and 3 to establish that, and explain how, the instrumental intelligence becomes general intelligence. However, this leads to complications about the plausibility of superintelligence: Müller and Cannon (2022) present an argument for thinking that general intelligence and orthogonality are inconsistent, and that a narrow Superintelligence is not an existential risk.

The existential risk of artificial intelligence, therefore, is heavily conditional on a definition of intelligence that sees it as instrumental and orthogonal to a final goal, and both notions which fall cleanly out of a "problem-solving" conception of mind, rooted in notions of computation and function.

Finally, alongside computation and functionalism, information also plays its part.

Recall the difference I made in §3.3.2. between "control" information – information for the purposes of coordination in a system – and "why" information – information offering understanding to the agent in question, from their perspective, of the meaning of their coordinated behaviour. In that section I tried to show that information, as a cognitivist, problem-solving notion, works with the same assumption, taking the cognitive situation as already-given, and that the notion of information in use is the "control" sense of information. This "control" sense of information works when the goal is already-given and the model has to describe *how* the agent achieves its goal. Just as with computation and functionalism, the control sense of information excludes an account of how the agent came to be in the situation that it is in, with the goal that it has. Control information excludes an account, from the perspective of the agent, *why* that agent is pursuing that goal.

One way of seeing it is in terms of the Orthogonality thesis again where "control" and "why" information are on "more or less independent" axes. The electrical communications machinery with which Shannon was concerned processed "control" information. Wiener's cybernetic systems too are

[50] See (Müller and Cannon 2022) for a detailed propositional reconstruction of the arguments that AI represents an xrisk. The paper is not, however, intended to identify or critique underlying cognitivist models or assumptions.

described with "control" information. In both cases "why" information is not part of the differences that are processed in the sense that such machines do not conduct cognitive work on why they should be pursuing their goal, purpose, or solving their problem. That much is already-given to them (and in any model of the process) and is thus not something they can or do in fact do.

I want to bring this chapter to a close and, at the same time, take some stock of where we are in the thesis because the next chapter on the post-cognitivist view of cognition brings in an altogether different perspective.

In order to do so, I want to present the notions of cognitivism and AI xrisk "in practice". I will first present Stuart Russell's example of how a robot tasked with even a banal goal such as "fetch coffee" can become an existential risk. Then I will stage a conversation between two positions. One the one side is a proponent who is familiar with the story and concepts involved in AI xrisk. The other character is an expression of the position of "common-sense" intuitions about intelligence, albeit, informed somewhat by the "already-given", "problem-solving" view of cognitivist assumptions involved in AI xrisk, such as I have claimed anyway.

The point of this next section is to summarise the thesis so far and show in the form of a conversation where the cognitivist assumptions do their work and, ultimately, that existing thinking about AI is distinctly cognitivist.

4.4. Fetching Coffee

Russell (2019) invites the reader to imagine a robot tasked with fetching a coffee, and the consequences of such an otherwise simple objective. Recall the discussion of instrumental goals from the second chapter on existential risk, superintelligence, and orthogonality - Russell claims that important "instrumental goals" fall out having the goal to fetch coffee such that the robot may come into rivalry with humans and then, if the robot is, or becomes, more intelligent than humans, that such a rivalry could be fatal for humans. He says that this is the case for more or less any goal that may be given:

> "Suppose a machine has the objective of fetching the coffee. If it is sufficiently intelligent, it will certainly understand that it will fail in its objective if it is switched off before completing its mission. Thus, the objective of fetching coffee creates, as a necessary subgoal, the objective of disabling the off-switch. The same is true for curing cancer or calculating the digits of pi. There's really not a lot you can do once you're dead, so we can expect AI systems to act pre-emptively to preserve their own existence, given more or less *any* definite objective… It is important to understand that self-preservation doesn't have to be any sort of built-in instinct or prime directive in machines…There is no need to build self-preservation in because it is an *instrumental goal* – a goal that is a useful subgoal of almost any original objective. Any entity that has a definite objective will automatically act as if it also has instrumental goals." (Russell 2019: 141)

The fundamental problem seems to be the failure on the part of the robot to appreciate what *we* would take to be relevant for the problem that *we* want it to solve. It shows up in the various hypothetical problems: Russell's "control problem" of how to *control* a machine that is more intelligent than humans

(2019); his "Gorilla problem", the question "of whether humans can maintain their supremacy and autonomy in a world that includes machines with substantially greater intelligence (Russell 2019: 132); and the "Value Alignment" problem (Soares 2016, Gabriel 2020, Cannon 2022), the question of how to ensure that AI shares human values such that it doesn't kill us, instrumentally or otherwise, when executing an objective.

Insofar as a capacity to appreciate what is relevant is a shared root of all these problems, the position I have been taking in this thesis is that this stems from taking the cognitive situation as already-given as the cognitivist models does. Following the discussions in this chapter, it is hopefully the clearer now that the kind of thinking Russell expresses in this example expresses distinctly cognitivist assumptions.

Here is a short dialogue between a prototypical and hypothetical cognitivist (Cog) and a sympathetic but bewildered philosopher who is trying to reconcile his stubborn common-sense with the paradoxes cognitivism presents (Sense).

Sense: So, you're telling me that the robot which is tasked to "fetch coffee" may become an existential threat…because of instrumental goals?

Cog: Yes. For example, if the system is intelligent, it will understand that it must stay alive in order to fetch the coffee, so will take rational precautions to stay alive. This is the instrumental goal of "self-preservation".

Sense: That seems a bit extreme. I'm not thinking about how to stay alive when I make my coffee. I can see that it is a rational inference, but it hardly seems a relevant one…

Cog: The definition of any given problem logically implies it. That said, the robot will only invest as much resources in that instrumental goal as is rational to do though. And more generally, Russell's "uncertainty" approach advocates making the robot uncertain about what our final goals are, and turning to us for evidence about what they might be. So according to our most developed theoretical approach, the robot won't necessarily kill us to stay alive in order to "fetch coffee".

Sense: Ok. It still seems odd that we have to control for an otherwise, and so-called, rational system potentially killing us. Yes, we all make inferences about what is instrumentally useful, but "self-preservation" never really presents itself as one to be worrying about. In fact, it seems rational *not* to spend time and energy dealing with endless scenarios that *might* happen. My body does a pretty good job of keeping me alive. If it gets to a point where staying alive is a problem, fetching coffee is not something I'm going to be thinking about, and if it *is* something I happen to be thinking about, I will at least be asking myself whether it is really worth it to fetch the coffee – but you say that *isn't* something such a robot would do?

Cog: Yes. Bostrom's Orthogonality Thesis means that robot's final goal and its capacity to achieve it are independent things. The robot is focused on achieving the goal, not thinking about whether the goal is worth it. That's what instrumental intelligence means, intelligence in service of "more or less any goal".

Sense: That sounds like computation too?

Cog: Yep, computation describes a process of cognition that is goal-independent.

Sense: So computation and instrumental intelligence are the same thing?

Cog: Well, they're definitely close. They're *about* the same thing, yes, but conceptually at least, they are distinct. We could say that computation describes cognition whilst intelligence is a *measurement* of cognition, in some admittedly unclear manner. Or, computation describes how a problem is solved, and instrumental intelligence is a measure of the system's capacity to solve it, or performance at solving it. They are both about the *function* of the mind, true.

Intelligence is also a more socialised concept in the sense that the significance of the term is of greater social consequence than that of computation or even cognition. For the dangerous ways in which measurements of intelligence have been employed to organise and coordinate society hierarchically, (for e.g. standardised testing in education), intelligence is a far more controversial notion than computation or cognition. Computation is "naïve" in this regard, and ontologically neutral. When we talk about intelligence, we are usually concerned with human affairs, but computation as a rule-bound processing of information applies to more or less any agent-system, human, alien, or artificial.

Sense: Ok, but computation and *instrumental* intelligence are the same thing then, functionally at least? Because both describe a process that is essentially "blind" with respect to where it's going…because caring about what I am doing is conceptually distinct from my capacity to achieve it, because solving the problem I have and discerning from own my perspective why it is relevant for me to solve are conceptually distinct - so sayeth the Orthogonality Thesis?

Cog: Yep, but the orthogonality thesis is not just a conceptual distinction. It is largely empirically verified by the artificial systems we build. We can and do build systems to achieve goals, and they do not need to understand why they are doing what they are doing in order to do it. In fact, it makes it a lot more complicated.

Sense: Fair enough. Pure problem-solving is distinct from discerning why that problem is worth solving, from the perspective of the agent in question. Why think that is good model for *human* minds though?

Cog: Well, we are not concerned with modelling *only* human minds. The space of possible minds is vast and we want a model that quantifies over the most possible space. We want principles of Mind and minds, not just human mind.

Sense: What makes you think this is actually a model of mind though, and not just a model of technological artefacts that is being *applied* to minds, a theory through which to make sense of minds, including some things and excluding others, based on what's relevant to the functioning of machines?

Cog: Sure, everyone appreciates that "all models are false, but some are useful".

Sense: What is this model actually useful for then? It seems like it is useful for modelling the behaviour of machines, machines which run our political economies and the infrastructure of our planetary communication systems, but these models fail to account for some otherwise very basic things. In taking the cognitive situation as already-given, the model eventually cannot account for what is relevant to me as an agent, (the frame problem), why I am here as a whole in the first place (the binding problem), and how and

why my expressions can mean anything to me (grounding problem), it seems that the model is not useful for a significant portion of human-like cognition.

It definitely describes something, but it seems to be missing important parts of the picture, as if it has isolated something very basic and very fundamental – and the connection between information, entropy, and computation *cannot but* leave one with a sense that it is something very profound indeed - but it nonetheless seems to be only a smaller arc of the larger circle of cognition. A "bit" of the whole.

Cog: What do you mean?

Sense: Well. It seems like Shannon identified an idea as profound as it is simple. His model of "communication" describes an information "transfer" of sorts, or a "reproduction" between two points of a system as he puts it, where the *quantity* of information is proportional to the "choice amongst possible alternatives" – if there were more alternatives, then the particular choice is more informative. But this kind of information only speaks to the *coordination and control* of the system in question. "Control" information, as Sloman puts it, is for a "control" system; Russell is concerned with a "control" problem, not a "meaning" problem. Like computation and instrumental intelligence, there is no accounting for why the machine is to be controlled and coordinated in that particular direction though – why point the tractor, rifle, car, in *that* direction; why give *that* particular problem to the computer?

Why those coordinations matter and are relevant to the agent in question is not included in the model of the functioning of the system. The "why" – why I am using the screwdriver to scratch my back or throwing it to play fetch with my dog – isn't accounted for. It is excluded from the off as a part of the "cognitive process".

It's no wonder they spoke of "control". I may not actually, ultimately be any more free, in a metaphysical sense, than a machine, but I make sense of the world in terms of what is relevant to me, which is to say, from a 1st person perspective. A model of my mind in which somebody else is defining from a 3rd person perspective what is meaningful to me is an obviously misguided model, and a little too close to political authoritarianism. Entire social justice movements like the waves of feminism seem to have been about women getting to define their own experience and sense of themselves as women, so this is hardly a trivial point.

Basically, it seems like these exclusions are problems for cognition *because* they're excluded. They are the rest of the circle that is assumed, but not included in an explanation of the curve of the arc.

Cog: Ok, so are you suggesting that if we include the rest of this "circle" in an account of cognition we can solve these problems?

Sense: I'm saying I think we won't have to. If the model includes the genesis of the cognitive situation from the beginning, then it seems like they never arise in the first place.

Sure, these exclusions and "already-givens" made sense in the time of Shannon and Turing and so on because it was obvious that we were talking about machines that *we* built for *our* purposes, which is to say, their purpose and assemblage was obviously explained by *our* goals. Why anything was relevant could only be specified by reference to something outside the system. That makes sense for machines and control systems, but

somehow this is now the picture we have for cognition in humans, and we have carried over these ideas and their exclusions, fitting ourselves into the mould of this model rather than moulding it to us. It is one thing to conceptually distinguish meaning and a capacity to pursue it as the orthogonality thesis does, but as a model of a machine has become a model of mind, it seems to have left meaning amputated in the process.

We needed the distinction then, and now it seems like we are trying to figure out how to put the parts together as a whole. But we don't have a way in the cognitivist paradigm of explaining why they would be integrated or whole – the binding problem! We even critique such suggestions as expressions of anthropocentrism and so paint ourselves into a corner of having to accept and rationalise something which obviously seems to be incomplete. Relevance is not a necessary or relevant thing to include in a model of the workings and coordinations of a computational system that has an already-given goal, but it *is* a relevant thing to include in a model of the human mind, at least if we want it to have a chance at accounting of the ways in which we experience meaning.

Cog: Alright. Because humans define the problems for machines to solve, including the rest of the arcs of the circle greatly complicates things though. An account of the problems for which we as humans build these systems would involve an account of sorts of entangled economic, social, political, psychological and who-knows-what sorts of reasons. Then it isn't a model of mind anymore, it's a model of societal complexes of production and, even assuming such a (2^{nd} order) model were possible, what use could it possibly have?

Sense: I don't know. But it seems like an important question. Let us treat it, not as the end of the inquiry, but the beginning.

4.5. Conclusion

In this chapter I tried to substantiate my claim that the narrative of AI existential risk is rooted in cognitivist conceptions of mind. I began by noting the underlying functionalism in the notion of "the space of possible minds", the anthropocentrism-avoiding inspiration behind much of the thinking about superintelligent AI. I then turned more explicitly to discuss how the AI xrisk narrative falls out of the cognitivist conceptions of computation, information, and functionalism. In particular, I noted the relationship between the conceptions of instrumental intelligence and computation, and how superintelligence is a logical extreme of these conceptions. Moreover, instrumental intelligence and computation are both goal-independent, which allows for the logical possibility of the Orthogonality Thesis. The Orthogonality thesis also reveals the functionalism at work for the way it decouples functions of mind that are otherwise taken to be whole. Finally, information plays its part too because cognitivism employs a conception of information that concerns control and coordination relative to an already-given goal, and not a conception of information that informs an agent with a sense of why their goal is meaningful to them.

I tried to summarise with a dialogue between a proponent of the position that AI is an existential risk, and a proponent of a "common sense" position who needs convincing.

So far in this thesis I have tried to set up in the following way my response to the question "can AI become more ethical than humans?". I began by situating the question in the context of the proposition that AI is an existential risk. Against that background, I pointed out that there are two paradigms of

mind research – cognitivism and post-cognitivism – and each has a different view, generally, on AI, and thus on the question of whether AI can become more ethical than humans. So far I have discussed cognitivism, having characterised it as a "problem-solving" model of cognition. I chose to present cognitivism first because, I claim, it is the way we are currently already thinking about AI. In this, the fourth chapter, I tried to lend some weight to this claim by pointing to the ways in which the AI xrisk narrative leverages the key cognitivist concepts of computation, information, and functionalism.

Now, in order to consider the existing alternative view on AI and the question of this thesis, I will turn to consider the post-cognitivist paradigm. In the next chapter I will present the fundamentals of the paradigm and its model of cognition through the enactive and "4E" theories of mind and cognition. I hope to bring out how, in the "enactivist" conception of mind, the genesis of the cognitive situation is an integral and paradigmatic feature of the model. This sets up a neat juxtaposition between the cognitivist and post-cognitivist paradigms of mind so that, going into the final chapter of the thesis, I am able to offer responses to the question of the thesis which consider what both paradigms might say.

5. Post-Cognitivism: Enactivism

> There are at least two kinds of games. One could be called finite, the other infinite. A finite game is played for the purpose of winning, an infinite game for the purpose of continuing play.
>
> - James P. Carse, *Finite and Infinite Games*

> Computers are useless. They can only give you answers.
>
> -Pablo Picasso

So far, I have presented the cognitivist paradigm of mind and suggested that its model of cognition can be characterised by its focus on "problem-solving". The key assumption of cognitivism which produces this view and model is that it takes the very cognitive situation as already-given – the agent, situated in the world, trying to solve their problem as they encounter it from their perspective, this basic set-up is taken for granted and is the starting position of the cognitivist model of cognition.

What we will see in the post-cognitivist model[51] I present in this chapter is a model in which the genesis of the cognitive situation is key to the story. In fact, for the cognitive process of the enactive and "4E" theories of cognition I will discuss here, the "autopoietic" individuation of the agent and the cognitive process itself are more or less one thing. That is, the process of an agent enacting itself *is* the cognitive process of that agent. The details of this claim and the processes involved are the content of this chapter.

The main distinction, vis-à-vis cognitivism, is that rather than a model of cognition starting from an already-given situation and moving through a process of problem-solving, the post-cognitivist model begins with the dynamics of how an agent emerges and regulates itself in the world as an individual, self-organising and self-maintaining in response to perturbations it encounters. Inspired by biological systems science and the dynamics of complex adaptive systems, in this model the basic demand on the individual is to make sense of what it encounters in order to stay alive. The "problem" is therefore not a fixed one amenable to solution and closure in the way that the already-given problems of cognitivism are. Rather, the "problem" is open-ended, continuously mutating and evolving as the individual moves in the world it perceives. In an important sense, there is no end-state because the demand is not a finite, well-defined, already-given thing. Playing with Carse's quote at the top of this chapter, cognition in the post-cognitivist model is an "infinite game" in which the purpose or "problem" is to keep the game going. This is not so much a matter of coming to a solution as adequately responding in the world. In order to do so, the demand on the individual is to perceive and make sense of its world where "success"

[51] I should reiterate a point I made in the introduction. "Cognitivism" is not always a common term in philosophy of AI, though it is in philosophy of cognitive science, where it is used mostly by post-cognitivists. Similarly, here, "post-cognitivism" is not such a common term or mode of self-identification either – proponents of the Enactivism and 4E cognition I discuss in this chapter are unlikely to refer to themselves as "post-cognitivists". I use the term to emphasise a basic difference between the models of the respective paradigms, namely, one sees cognition as what happens once the cognitive situation (agent with a problem in an environment) is taken for granted, (cognitivism), while the other (post-cognitivism) sees cognition as involving the emergence of the cognitive situation itself.

in this regard is the open-ended continuation of the process, more or less regardless of the direction it takes.

It is this phenomenon that has led me to characterise the post-cognitivist paradigm as "problem-defining". In the first section of this chapter, I try to articulate in more detail what I mean by "problem-solving." The characterisation is, I think, a reliable conception of the models I describe in this chapter, but I chose it too for the way in which it sets up the problem-solving – problem-defining juxtaposition I am putting to work in this thesis. It is important not to get caught in the trap of thinking that "problem-defining" is the new problem to solve. Whilst it is not entirely false, it would be a distinctly cognitivist way of viewing the phenomena post-cognitivism picks out.

Post-cognitivism has its own concepts, language, and way of speaking about all this and it is important to move with that language. The language of cognitivism, with its problem-solving goal-orientation, can bias an encounter with post-cognitivism. This is something Maturana himself[52] notes already back in 1973 in *Autopoiesis and Cognition: The Realization of the Living* when, concerning the capacity of 'autopoiesis' to fully characterise the organisation of living systems, he says "notions of purpose, function, or goal are unnecessary and misleading" (1973: xix)[53]. It is important to bear in mind because it is entirely possible to see and describe the post-cognitivist model in cognitivist terms. The risk is that of missing the point, or failing to appreciate the distinct things the post-cognitivist theories have to offer to our understanding of cognition. In section 5.3. for example, I consider some of the "4E's" of cognition and how they seem, so-far, to have been understood in a remarkably cognitivist problem-solving fashion. Therefore, in this chapter, I have tried to present the ideas in their own terms, minimising "translation". The language *is* different, as is the style. Cognitivism lends itself to analytic philosophy in both its manner of expression and its mechanical way of thinking – a phenomenon which deserves further scrutiny – whilst post-cognitivism, being inspired on one side by phenomenologists like Merleau-Ponty (Varela et al. 1991/2016, Thompson 2007) and eastern spiritual insight practices on the other, finds itself expressed in language that often asks of the reader to meet it halfway. I have tried in this presentation of what I refer to as post-cognitivism to be faithful to the way in which the paradigm expresses itself.

The goal is this chapter is thus to present the main paradigmatic alternative to cognitivism. The different view of mind and cognition it offers amount to thinking differently about artificial intelligence and its potential – generally, but specifically as it concerns the question of this thesis. To this end, in this chapter I present the Enactive and 4E theories of cognition and characterise it, by comparison to the cognitivist model, as a "problem-defining" paradigm of cognition. Once this is set in place, I can address in the next chapter the question of whether AI can become more ethical than humans from the perspective of both paradigms.

The chapter on cognitivism focused on some of its key, paradigmatic concepts and suggested that computation, information, and functionalism support its problem-solving orientation. This chapter on post-cognitivism will similarly focus on specific concepts. In the same way that I presented the cognitivist concepts with a view to showing how they take the cognitive situation as already-given, here I will present the enactive and 4E theories of cognition with a view to emphasising how these theories describe the genesis of the cognitive situation. The point is not that post-cognitivism accounts for the cognitive situation and cognitivism does not, but what happens when the genesis of the cognitive

[52] If Francisco Varela is the father of the enactive paradigm, then Humberto Maturana, as his early mentor and then collaborator, might well be the grandfather.
[53] More specifically, the quote comes from Maturana's introduction to his 1970 "Biology of Cognition" paper, which he has expanded for *Autopoiesis and Cognition*.

situation is *included* in the model of cognition. Principally, it generates a view and model of cognition in which the agent navigates the world it perceives on its own terms, including, defining the features of its world, and the value thereof for and by means of itself. It is not within the scope of this thesis to explore how, in this way, post-cognitivism never creates for itself the frame, binding, or grounding problems the way cognitivism[54], but what *is* important is that the place and potential of AI looks altogether different from the perspective of the post-cognitivist model I outline here, and from and the enactive theory in particular.

The enactive theory of cognition can be viewed as a piece of the 4E theory of cognition, the theory that cognition is embodied, embedded, extended, and enactive (Aizawa 2018, Newen et al. 2018). The enactive theory (Thompson 2007) is perhaps the most potent of the combined E's for bringing out the "problem-defining" character of post-cognitivism and for this reason I give it the most weight in this chapter. The dynamics of enaction are key to understanding how the cognitive situation emerges. The enactive theory offers a view of how an agent comes to be in the world – *its world,* according to what it is capable of perceiving – with the particular demands and "problems" it discerns and defines. The final section of the chapter is devoted to the remaining E's of 4E cognition.

As in the previous chapter, I am not presenting and arguing for a definitive account of either the Enactive theory of cognition or 4E cognition, let alone a definitive account of post-cognitivism. There exist detailed, book-length accounts which explore Enactivism and 4E cognition in its most up-to-date developments and details (Varela et al. 1991, Thompson 2007, Newen et al. 2018). Here the aim is to present a characterisation and understanding of the paradigm which is both accurate and faithful to it, and which highlights those features most relevant to question of this thesis.

The characterisation of post-cognitivism in terms of "problem-defining" can be misleading in some important ways, so before getting into the details of the theories of cognition, I devote the first section of this chapter to clarifying what I mean by the notion. "Problem-solving" is comparatively straight-forward, but there are more nuances to "problem-defining".

Having hopefully established a little more clarity about the notion, I then dive into a presentation of the enactive theory of cognition, focusing on four concepts, with a view to substantiating the problem-defining characterisation.

The first two concepts of enactivism are entangled. The *life-mind continuity thesis* claims that life and mind are continuous – where there is life, there is mind[55], and vice versa (Thompson 2007: 128, Kirchhoff and Froese 2017). This plants the paradigm's research in the context of the biological and ecological sciences of life. This detail is important for making sense of the other of the concepts. The second concept, *autopoiesis*, identifies a minimal condition for life, and, thus, in this picture, mind. The process and condition of autopoiesis is one in which organisms generate the very conditions for their own existence: auto = self and poiesis = creation. The third concept is the conception of cognition for living, autopoietic entities – *sensemaking.* Sensemaking involves a capacity of discernment in which defining things, with a sensibility and sensitivity to what is relevant to the survival of organism, is central. The last concept ties things together. The *dynamic co-emergence of self and world* describes a

[54] In other work, (Cannon 2022), I make a small step in this direction, specifically as regards the frame problem, and in the context of the Value Alignment problem of AI.

[55] The life-mind thesis refers specifically to "mind". I have been using "cognition" to avoid the question of consciousness that discussions specifically of "mind" tend to invoke. In my understanding, the life-mind thesis does indeed refer to "mind", but questions of consciousness and the like are not my aim here, so I am using the two terms interchangeably for now unless otherwise stated.

transcendental dynamic in which self-and world co-emerge together: as the autopoietic individual enacts itself, it "brings forth" a world, the world it is capable of perceiving.

The aim then here is to offer a focus on problem-defining as a meaningful frame of reference through which to make sense of the post-cognitivist paradigm. This establishes a neat complementarity between the problem-solving of cognitivism and the problem-defining of post-cognitivism. It will be in terms of this distinction that I move on to the final thesis chapter to consider whether AI can be more ethical than humans. Throughout this chapter, it should be noted that post-cognitivism is not only concerned with problem-defining as I have defined it; my claim is that post-cognitivism is concerned with both problem-defining and problem-solving.

5.1. Defining Problem-Defining

Art, like morality, consists in drawing a line someplace.

- G. K. Chesterton

It should be restated at the outset here that "problem-defining" is not terminology of post-cognitivist theories, though the underlying idea is present in descriptions made by some post-cognitivist authors. It is a term of art in this thesis. The point is to bring awareness to a basic but distinguishing feature of the post-cognitive model of cognition, vis-à-vis cognitivism's model, namely, the significance of the dynamics which generate the cognitive situation in the first place.

Now, authors on the post-cognitivist side of things do not talk in terms of "problem-defining". The local and formal terminology is "sensemaking". Everything that has been said about problem-defining is said more formally of sensemaking. Sensemaking includes far more than just "problem-defining" though because it includes the process of individuation and development of the agent in question. That process is vital to why an agent defines and individuates things as they do. The individuation of the cogniser is taken as already-given in the cognitivist paradigm – we speak of "an" AI system without having clearly defined consensus about how to individuate these systems[56].

I could have referred to the two paradigmatic models here by their respective technical terms for cognition, computation (for cognitivism) and sensemaking (post-cognitivism's term), but these terms don't distinguish the paradigms from each other with any clarity unless you happen to already be familiar with both models. "Problem-solving" and "problem-defining" neatly present the distinction of the paradigms. Relative to 'sensemaking', the conception of cognition of the Enactive account I present in this chapter, problem-defining concerns the broader way in which the paradigm appears to think about cognition and model it.

In this regard, at the risk of oversimplification, problem-defining takes place before problem-solving. It is the part of the cognitive process that is "already-given" in the cognitivist model. Again, an account of problem-solving excludes an explanation of how the agent comes to be trying to solve the particular

[56] In *Atlas of AI*, Kate Crawford points out that there are different ways and places to draw the line. She argues that a "map" of AI which takes into consideration the means of production of contemporary AI technologies reveal lines of geopolitics, socioeconomics, colonial histories, technoscience gender relations, modern labour conditions, and more. The strongest way to put the point here is that a model of AI which takes the individuation of the system as already-given, excludes from our understanding of that system all the above dynamics and the effects of taking them into consideration.

problem that they are because the problem is already-given. One of the effects of this exclusion in a model of cognition, so I claimed in the previous chapter, is that we are led logically to the proposition of a superintelligent AI that is ambivalent about the existence of humanity unless it makes a difference to its capacity to achieve its goal.

"Problem-defining" concerns the genesis of the cognitive situation and one of the effects of *this* line of thought is a theory of agents which are inherently and continuously reevaluating what they encounter, not relative to a "final goal", but in open-ended fashion. They do not exist in order to solve a problem. They just exist and more or less try to stay that way[57], relative to whatever they encounter.

With this view in mind, this section works up to a conception of problem-defining as the process of an individual making specific, sense of their open-ended situation, or, making sense specific to their situation. This involves two things, firstly *orienting* themselves, and then, given that orientation, *discerning the meaning of encountered features*. As a preliminary definition of problem-defining therefore, we can define it as *an individual's endogenous orienting discernment of situation and meaning.*

Ultimately, problem-defining is an everyday language way of talking about "sensemaking", the Enactive Theory's concept and account of cognition. As a reminder, in this thesis the enactive theory of cognition plays the same role for post-cognitivism as the classical computational theory of cognition played in the previous chapter for cognitivism. They are both representatives of their respective paradigms. For a technical and rigorous discussion of problem-defining, then, look to the notion of "sensemaking". As far as I am concerned, "problem-defining" is more colloquial means of expressing this more sophisticated concept. (Remember too that post-cognitivism's model of cognition is not exclusively about "problem-defining". It allows for "problem-solving too.)

The importance of "problem-defining" is therefore the following. So far in the thesis, it has been asserted that cognitivism's model of cognition takes the basic cognitive situation as "already-given", the cognitive situation being the situation of an individual with a specific problem to solve in a particular environment. "Problem-defining" is an everyday language way of recognising that something must take place in order to get to that situation to begin with. Unlike much of standardised Western schooling, and even university-level, education, not to mention psychometric tests, the demands on human cognition are not exclusively pre-defined test questions. At its most basic, in order for the agent to take itself to have a problem to solve, it must first discern and define a problem.

A basic etymological analysis of the verb "to define" offers something of value. To "De-fine": to/of – finitude; a rendering-finite, and by implication, a rendering-finite of that which is not finite, or, at least is less so. Insofar as we can grasp the finite, but not the infinite, de-finition affords grasp, grip, and traction. Well-defined problems admit of easy solutions, proportional to the purchase on the problem. Standardised test questions are in this sense easier to solve than problems that are undefined and require work to figure out what to do. To a first approximation, this work is what I mean by problem-defining.

Problem-defining can be misleading. As a verb, it suggests it is an activity or active process on the part of an agent. What I am pointing at is closer to Heidegger's famous notion of "thrownness", (Geworfenheit), which points to the way in which we existentially find ourselves always-already in the world (Heidegger 1962: 174). Like the phenomenology of thrownness, problem-defining is more a

[57] Obvious counterexamples might include cases in which an organism actively wants to end its life. Problem-defining is still happening in these cases, though. Some might say then that the problem-defining capacities are failing, but I don't think that is quite right. As I go on to explain, I think we could say in such cases that the individuals are still problem-defining and, simply, either their discernment or basis of discernment leads to senses and experiences in which ending their life is something that makes sense to them.

phenomenon of finding oneself always-already in a situation. When we find ourselves in awareness[58], the cognitive demand is to makes sense of what we perceive in a way that keeps us alive.

The "thrownness" quality of problem-defining does not mean that the problem is already defined for us. To the contrary, problem-defining concerns establishing an orienting discernment in this situation in order to make sense of what our situation means for us.

We can talk about orientation and meaning specifically. Orientation is the basic precondition for being able to make sense of a situation (Stegmaier 2019)[59]. If problem-solving assumes that the problem is known, problem-*defining* may involve making sense of the unknown; Stegmaier says something interesting in this regard: "In orientation, one is at first dealing with something one does not yet know about: a new situation." (ibid: 1) The case of geographic, land-based orientation is an easy case for intuitions. If I do not know or cannot tell which way is north, or any other cardinal direction, I cannot begin to establish where I am. If I do not know where I am, I cannot tell if I am in the right or wrong place and thus cannot make much sense of what my situation is.

There is a dialogue in *Alice's Adventures in Wonderland* which cleverly illustrates the significance and necessity of a basis of orientation in problem-defining. Whilst newly exploring the paths of Wonderland, Alice asks the Cheshire Cat for directions: the Cat responds that it depends where she, (Alice), wants to get to. Because Alice is new to Wonderland, she has no idea where she is, so she has no idea where she even ought to be wanting to go. She can't tell what her problem is and is more than simply lost because she doesn't even have a basis on which to determine whether she is or isn't where she is supposed to be or where she should want to go. She has no basis for orientation. Thus, everywhere is as meaningless to her as everywhere else. In a state of frustration, she says she doesn't much care where she goes, as long as she gets *somewhere*, to which the Cheshire Cat responds that she is sure to get there, as long as she walks long enough. In a catch-22, she wants to get to a "*where*" – a well-defined location that can be a basis of navigation, one which affords for her a sense of what the "problem" actually is for her, but in order to do that, she needs a basis of orientation in the first place.

Orientation is therefore a necessary basis for determining the meaning and relevance of features we encounter – the condition of possibility for making sense of our situation and thus "defining the problem". Without a basis of orientation, the meaning of the features of my environment and situation is unclear. Stegmaier, echoing Heidegger's "thrownness", asserts that "at first, one does not search, but one finds" (2019: 5). This is phenomenological kind of thinking is foundational to the enactivist theory I discuss in this chapter. Whilst it is the predominantly the work of Merleau-Ponty from which the enactive theory of mind takes phenomenological inspiration, Heidegger's work is also a rich vein for the model; in *Mind in Life*, Evan Thompson, citing Heidegger, speaks to precisely my point here:

> "'A stone never finds itself but is simply on hand. A very primitive unicellular form of life, on the contrary, will already find itself…' (Heidegger 1962: 255) This notion of 'finding itself'…(*Befindlichkeit*) defines the phenomenological meaning of being-in-

[58] In speaking of "awareness", I do not mean to necessarily invoke phenomenal consciousness. Something akin to "perceiving", or a capacity of perception of some means that need not involve consciousness, such that individuals to which we might be reluctant to attribute consciousness, like single-celled organisms or machines, are included here.

[59] As a concept, orientation receives on the whole remarkably little philosophical attention in philosophical work on mind and cognition, despite having much to offer. See Werner Stegmaier's *What is Orientation?* for a book-length discussion inspired by a short work of Kant's: *What Does it Mean to Orient Oneself in Thinking* (Stegmaier 2019: 3).

> the-world (see also p.165-166, where Heidegger explains 'in-being' as analogous to the being proper to life, by contrast with non-living things.)" (Thompson 2007: 455, ft. 11)

The distinction between searching and finding is insightful here if understood in the sense that searching for something seems to imply already knowing what we're looking for, while finding something seems to imply receiving something unexpectedly, which is to say, not knowing. With problem-defining, we do not assume that a problem exists independent of the agent, waiting the agent's discernment thereof. We can say that we are not searching for a problem so much as finding it, which is to say, finding ourselves in a particular situation.

It should be made clear that this does not mean "finding ourselves to have certain *goals*". This would be an example of a cognitivist reading. The establishing of orientation in a situation is open-ended. I do not need to have goals in order to be oriented in the world, though of course having goals can indeed afford orientation.

Returning to the geographical example, if I am disoriented such that I can't tell which way is north, the meaning of the ridgelines, peaks, and valleys in which I am embedded is unclear. Recall that "meaning" is being used throughout this thesis as the particular significance that features of our situation take on for us, the particular difference that they make to us. By virtue of the significance these features take on for me, they can be more or less relevant to me. With or without a basis of orientation, a pebble by my foot is going to be less relevant than major features like a river or ridgeline. The particular significance that the ridgeline and river take on in the situation for me though, will now depend on my established geographic orientation. It may be something like "the ridgeline is on my west but should be on my east".

It is only once these processes of "orientation and discernment of meaning" have occurred that problem-solving comes into play – "I need to get to the other side of this ridge". Solving that problem is only possible once orientation is established because it creates a condition in which features can take on a specific meaning relative to the individual in their situation.

It is important to note that problem-defining as orientation and discernment is not necessarily a factual or objective process in which, after a certain amount of effort, the right and real problem is eventually discovered. Problem-defining does not suppose any kind of objectivity or realism because it is not primarily about discerning objects, it is, again, about establishing the possibility of orientation and meaning from the perspective of the agent[60]. This is more or less clear in the context of geographical navigation, but in other contexts the notion of "problem" may be misleading. Let us consider the example of problem-defining the "problem" of artificial intelligence.

Given AI is the project not of an individual but a whole collective of humans, we can say that, collectively, we are currently engaging in problem-defining with regard to artificial intelligence. That is, we are engaging in a process of orienting discernment of the meaning of AI, the particular significance it endogenously takes on for us. Whilst we have identified more and less well-defined problems, or problem domains, like AI ethics and ethics of AI, we are continuously having to reorient ourselves as AI disrupts our orientations by doing things which were once thought to be uniquely or

[60] The realism question is a nuanced one for any post-cognitivist theory. For the Enactive Theory of mind discussed in this chapter, and which serves as the expression and representative of "problem-defining" post-cognitivism in this thesis, it remains to be fully understood what the model of cognition means for realism questions. See §5.2.4. on Enactive Dynamic Co-Emergence. See also the quote from (Thompson 2007) on the following page.

distinctly human, such as producing art and music, performing economic labour, stimulating preferences with recommendation systems, and so on. These technological achievements are forcing and/or inviting a reconsideration of the meaning of human activity and challenging our sense of orientation. Whilst we have defined some specific, clearer problems to solve with AI, we continue to work to figure out what the "problem" is. This thesis is part of the problem-defining efforts.

Now, though we would want to say that many problems which we define vis-à-vis AI are entirely real, we would not expect to finally arrive at *the* problem of artificial intelligence. It would certainly be interesting, but also controversial, to suppose as much. Instead, as artificial intelligence continuously disrupts our orientation about what, if anything, is special about human beings, by continuously proving capable of performing (well-defined, already-given) tasks and labour, the meaning of human creativity, labour, embodiment, and relationships are also being transformed, and so we are continuously trying to find out what these things now mean for us. So, again, problem-defining in this context does not mean arriving at a well-defined conception of the problem, but establishing orientation and meaning in the situation into which we are "thrown", which means labouring under uncertainty. Stegmaier points out that "it is the basic condition of all orientation to operate under uncertainty" (2019: 10).

This example of artificial intelligence concerns society and collective problem-defining. For an example more clearly involving *individual* cognition, we can consider basic cases of individual human perception. If problem-defining is the process of orienting discernment of meaning, consider examples where perception can be disoriented. If we return once more to the example of geographical navigation, if there is heavy fog, it may be impossible to make out any features of a situation at all, as occurs with "white-outs" in snowy mountain conditions. In such cases, "up" and "down" are about as oriented as we can be. We can lose this orientation underwater, too. Divers are taught in such circumstances to release a small amount of air as bubbles as pressed upwards by the water pressure toward the surface, indicating the direction of the surface - "up". In both cases, an individual can be in the same situation as Alice is in Wonderland. Obviously in each case the "problem" is in some basic and important sense simple, but in each case orienting discernment of meaning, aka "problem-defining", precedes any "problem-solving" that can be said to happen.

I want to discuss one last example. The case of "chemotaxis" can show that problem-defining applies to even minimally cognitive individuals. Single cells engaged in chemotaxis move in response to stimuli they encounter in a way that can produce behaviour even as sophisticated as navigating a maze (Tweedy et al. 2020, Weijer 2020). Such an organism can be said to be "problem-defining" insofar as it is engaging in orienting discernment of the meaning of its environment. It is recognised that these organisms follow gradients of chemical concentrations, either in the direction of nutritional sources like glucose, or away from sources of toxins, and that this behaviour is sufficient to navigate a maze structure (ibid). The glucose or toxin concentrations take on a particular meaning for the organism and, upon encountering a physical structure preventing mobility in a given direction, the organism uses the gradient as a basis of orientation and moves around the wall according to the gradient. This problem-defining does not involve coming to a well-defined conception of a problem, it remains a matter of orientation and meaning. Evan Thompson, an author central to the enactive paradigm I discuss in this chapter, puts it like this:

> "…although sucrose is a real and present condition of the physiochemical environment, its status as food is not. That sucrose is a nutrient is not intrinsic to the status of the sucrose molecule; it is, rather, a relational feature, linked to the bacterium's

> metabolism. Sucrose has significance or value as food, but only in milieu that the organism itself brings into existence." (Thompson 2007: 158)

Now, normally, this problem-defining as "orientation and discernment of meaning" is going to be adaptive in the sense that, if an individual cannot make sense of its situation in a way that is beneficial in the sense of keeping it alive, then we would want to say that they are not very good at "problem-defining".

In the examples above, if humans collectively define the "problem" of AI in a way that kills all humans, for example by developing artificial Superintelligence, we would want to say we have not done very well in terms of orientation and discernment of the dangers of AI. Similarly for the case of an individual who is geographically lost. If they cannot find their way out, we can say again that they're probably not very good as "problem-defining", though, of course, it may just be an extremely challenging situation in which even an experienced "problem-definer" in geographical navigation might fail. In the case of chemotaxis, if they cannot discern the difference between toxin and nutrition, it also seems right to say that the organism is not good at "problem-defining". The point then is that problem-defining is inherently related to what is beneficial for the individual from their own, endogenous perspective. This is something that is discussed in more detail in the sections of this chapter on Autopoiesis and Sensemaking.

An individual can be "wrong" in their problem-defining. There can be a mismatch between an individual's orientation and discernment of meaning and what we would want to say is actually beneficial for them. One established example of this is the notion of "evolutionary mismatch" from evolutionary theory. Evolutionary mismatch occurs when traits which were adaptive in one environment become maladaptive as the environment changes (Lloyd et al. 2011, Gluckman et al. 2019). A common example is the evolved trait in humans of preference for calorie-dense food, e.g. sugar, a trait selected in the calorie-scarce environments of pre-industrial environments, and which is now maladaptive in industrialised environments in which there is an abundance of calorie-dense, sugary foods. Our discernment that a bit of sugar is "good" would, in this sense, now be "wrong", but the quotation marks are there to emphasise the contingency of these discernments. They need not suppose a realism in any sense about the content of "right" and "wrong".

An example of collective problem-defining going "wrong" is smoking in the sense that it was once discerned to be beneficial and now is recognised to be a serious health risk, i.e. maladaptive. In the cases of individual discernment, cases of altered states of consciousness seems to involve a higher risk of "mismatch" too. Being drunk, for example, we seem prone to make maladaptive discernments, or under states of psychological duress, fatigue and sleep deprivation. Whether mismatch in problem-defining occurs in the case of chemotaxis and other microfauna is less clear. Unregulated growth of parasites or cancerous cells could be considered poor discernment in the cases that it kills the host and thereby ending the condition of possibility of parasite or cancer. Whether a chemotaxis can be wrong in their problem-defining discernment is unclear, but it seems plausible at least in principle that in an environment unfamiliar to it, there could be a mismatch between the chemotaxis' problem-defining, and what in fact was adaptive for it.

Problem-defining can also go "wrong" in "positive" ways, which is to say, way that are not necessarily maladaptive. For example, sometimes my dog seems to discern a threat in a tree stump that vaguely resembles a human, or discerns that my 3 year-old nephew is a threat when he only wants to play with her. This mismatch is "positive" in the sense that, in discerning danger where in fact there appears to

be none, she is "wrong" about something which might be good for her. There seem to be many conceivable examples of things we discern to be "bad" or maladaptive which are, or may be, good for us.

Summarising, problem-defining, then, is the endogenous process of an individual's orientation and discernment of meaning, and all happens before any problem-solving, where "meaning" is the significance specific to the individual in question.

It is now possible to distinguish the problem-defining character of the post-cognitivism model to come in this chapter, from the problem-solving character of cognitivism in the previous chapter. In order for there to be a situation of an agent or individual with a problem to solve, the problem (from the perspective of the agent) has to be defined (by that agent). This step is "already-given" in the cognitivist model. Moreover, for the problem to be meaningful *to the agent* in question, that is, for the problem to take on the particular meaning *for the agent* that it does, it must be the agent doing the discerning. Recall from the previous chapter, the systems under discussion, from Turing and his eponymous machines, to Shannon and his information theory, Wiener and his cybernetic systems, and Marr, with his Tri-level model of information-processing systems, all the systems were viewed and understood from a 3rd person perspective, which is to say, from outside the system. The problem-defining necessary for determining what problem the systems were to solve, was work conducted by a human outside of the system. The meaning of the features processed by the system were imposed by the engineer for the engineer's problems[61].

This concludes the discussion on "problem-defining". The key line of thinking to take away is this. Where problem-solving cognitivism takes the cognitive situation as already-given, the problem-defining of the post-cognitivist model points to what happens in getting into the cognitive situation. It involves an agent finding itself "thrown" in the world of which it is aware, and then required to establish orientation, telling "up" from "down", such that it can make sense and meaning, on its own terms, of what it perceives, all in continuous, open-ended fashion, in order to keep the infinite game going.

Finally, this conception of problem-defining should be held loosely and as a heuristic with which to make sense of the post-cognitivist model I describe in the rest of the chapter. As a reminder, the argument of this thesis is that what we think about AI depends on our model of mind, and given we have at least two fundamentally different models, we are going to at least two different pictures of AI, so for the question of this thesis, we are going to have at least two different answers. Post-cognitivism has a different view, and "problem-defining" is a way of characterising it.

5.2. The Enactive Theory of Cognition

Whilst the Enactive Theory of Cognition may be seen as the post-cognitivist counterpart to cognitivism's computation, there is good reason to think that it represents not just a theory, but a paradigm itself. The title of a jointly edited book by Stewart et al. (2010) on Enaction presents it thusly: *Enaction: Toward a New Paradigm for Cognitive Science*. The "toward" may be more or less significant. That is, Enaction may be the opening of something larger, like "4E" cognition (Newen et al. 2018) or it may indeed itself be the new paradigm. Because this is yet unsettled, in this thesis it is treated as a representative of the key elements post-cognitivism I want to highlight, without being

[61] This point is discussed in more detail in sections 5.2.2. on Autopoiesis, and is emphasised in section 5.2.2.3. on Internal Norms of Self-Regulation.

necessarily exhaustive of post-cognitivist thinking. For my purposes, it gets to the fundamental distinction with cognitivism, the problem-solving – problem-defining distinction.

The enactive theory of cognition has four key concepts: Life-mind continuity, autopoiesis, sensemaking, and dynamic co-emergence (Thompson 2007). These concepts articulate a condition of being in which discerning what is going on from its perspective is the condition of cognition. It is a condition of perspective in a frame of reference in which there is sensitivity to what is relevant, for the individual in question and from their perspective.

What I hope to bring out of this section is first and foremost the way in which, in the post-cognitivist paradigm, defining a problem, and a capacity to do so, is fundamental to an understanding of cognition and agents. Again, the long game here is that it matters what ideas we use to consider other ideas, and we go very different directions vis-à-vis thinking about AI becoming more ethical than humans with cognitivist or post-cognitivist ideas. These four concepts express the basics of a paradigm of cognition in which the determinacy of objects "out there" is not assumed, and rather the process in which they come to be defined for and by an individual, is fundamental (Thompson 2007, Stewart et al. 2010, Newen et al. 2018). For an excellent, condensed summary of the enactive theory of mind, presented in terms of design principles for "Enactive AI", Thomas Froese and Thomas Ziemke's 2009 paper "Enactive Artificial Intelligence: Investigating the systemic organisation of life and mind" deserves reading. Now we can turn to the enactive theory of cognition. The first concept to understand is the life-mind continuity thesis.

5.2.1. Life-Mind Continuity

The "life-mind continuity thesis" (Thompson 2007: 128-129: Kirchhoff and Froese 2017) contextualises the other concepts in this chapter, situating and embedding them in biological approaches to mind and cognition. Biologists Humberto Maturana and Francisco Varela are largely credited for inspiring mind research in this direction, away from cognitivist assumptions (Maturana and Varela 1987, Thompson 2007).

The idea is that where there is life, there is mind, and vice versa. Thompson writes, "According to this thesis, life and mind share a set of basic organisational properties, and the organisational properties distinctive of mind are an enriched version of those fundamental to life." (Thompson 2007: 128) There are stronger and weaker versions of the thesis. In his textbook introduction to the philosophy of cognitive science, Andy Clark says of the life mind continuity thesis:

> "the thesis of strong continuity would be true if, for example, the basic concepts needed to understand the organisation of life turned out to be self-organisation, collective dynamics, circular causal processes, autopoiesis, etc., and if *those very same concepts and constructs* turned out to be central to a proper scientific understanding of mind" (Clark 2001: 118)

Again, the significance of the life-mind continuity thesis is that what is relevant for a mind is defined by its condition as a living organism. This in turn is significant because it places biological and ecological principles at the centre of enactive understanding and exploring the nature of mind (Maturana 1970; Varela et al. 1974, Varela 1979, 1997, Maturana and Varela 1998; Thompson 2007). The second

concept, autopoiesis, is a firmly biological concept. Understanding the biologically inspired context of the enactive approach to mind is important for making sense how autopoiesis comes to have the place that it does in enactive approaches to mind and cognition.

There is a little more to say here about what more generally a biologically inspired approach to mind affords. The most important distinction it accounts for is the difference between the "goal-directedness" of the living and the artificial (Froese and Ziemke 2009: §3).

It is worth acknowledging that it might be disputed that a distinction between living and artificial begs the very question which AI research is exploring, namely, whether such a distinction is a difference that makes a difference. Froese and Ziemke use the word "artefacts" (2009: 19) to describe non-living systems, and one might very well use terms like "machine" or even "tool". The technologies which come under the umbrella of AI are interesting largely for the way in which they transgress the boundary between being mere "tools" without agency, and being something else, something with a sense of "agency" we usually reserve for a class of entities we consent to call "living". Recall from the previous chapter that speaking in terms of information-processing systems flattens this difference, or at least suggests that it is an uninformative difference insofar as living and non-living systems can realise the same functions. In any case, however one wants to name that which the line differentiates, AI is not an interesting enterprise unless one has a sense, implicit or explicit, that there *is* such a line which AI is blurring. So, whilst it might beg the question, it is a distinction which is more or less performatively assumed as the basis of AI being an interesting philosophical project. Failing that, for the time being, the onus seems to be on those who would deny the existence of a distinction to explain the differences we observe between living systems and non-living, or technological, systems.

With that caveat, there is an important distinction between the goal-directedness of the artificial and the living which biological thinking about mind affords:

> "…while (current) artefacts can only be characterised in this [goal-directed] way because of their involvement in a purposeful context that is external to them (extrinsic teleology), organisms appear to have the peculiar capacity to enact and follow their own goals (intrinsic teleology)." (Froese and Ziemke 2009: 19)

Another way of putting this, without the sometimes-controversial language of teleology, and in the language of this thesis, is that the "goal" in question of artificial systems is defined for it, exogenously and external to it. This is often the "function" of the machine, and terms in which the difference between function and malfunction is drawn. On the other hand, the "goal" of living systems is endogenously defined. This means that what is meaningful is defined *by the living system itself, for itself*, but in the case of the machine is defined externally for a machine. I do want to be careful here though. I stressed in the introduction to this chapter that speaking in terms of "goals" risks interpreting what is going in a cognitivist way. As we discuss the next concepts, the way in which the endogenous dynamics are open-ended – i.e. not necessarily goal-oriented – is something I want to bring out.

With that caution in mind, there is nonetheless an important line of thinking worth continuing here. There is an oft-cited passage in Kant's *Critique of Judgement* on the notions of "natural purpose" and "natural ends" which together articulate this difference of the goal-directedness in a way that sets up for the discussion of autopoiesis to come.

In §64, "On the Character Peculiar to Things Considered as Natural Purposes", Kant suggests two criteria of [living] systems understood as "natural purposes". He does not speak explicitly of "living" systems, only "systems understood as "natural purposes"". The "living" is an addition, but the line he is drawing is the same line, something recognised by Thompson (2007: 129-140) and Kauffman (1993: 4, 2019: 8-9)

> "Now in order for a thing to be a natural purpose, it must meet two requirements. First, the possibility of its parts (as concerns both their existence and their form) must depend on their relation to the whole… the second requirement is that the parts of the thing combine into a unity of a whole because *they are reciprocally cause and effect of their form*." (Kant 2007: 252) (emphasis mine)

Another way of putting this is to say that, in order for a thing to be a natural purpose, the parts are not arbitrarily assembled according to exogenous ideas and goals, (leading to the binding problem). Instead, the parts create a whole which reciprocally creates the parts. The part-whole relation here is not one of mere aggregation in which the whole is the sum of the parts. The parts are both cause and effect of their form as an integrated whole. This is an abstract articulation of the self-organising dynamics of autopoiesis, the next concept in this chapter.

For the moment, what matters is that "natural purposes" of living systems express this organisational character. In the next section of his *Critique of Judgement*, Kant talks about "natural ends". I have quoted it at length because he links his notion of natural purpose to "natural ends" with the example of a watch, in a way that brings out quite clearly the difference between the exogenous goal-directedness of a mechanical artefact like a watch, and the endogenous, "natural purpose" of a living organism. This is the same watch that was briefly mentioned in the discussion of the binding problem in the cognitivism chapter[62].

What Kant has to say in this example of the watch offers a sophisticated distinction between biology and mechanical artefact and is therefore worth acknowledging as esteemed intellectual heritage in this context.

In his discussion of "natural ends" in section §65, Kant makes specific reference to the role of self-organisation again as that which distinguishes living "natural ends" from machines:

> "In such a product of nature[,] every part, as existing through all the other parts, is also thought as existing for the sake of the others and that of the whole, i.e. as a tool; … an organ bringing forth other parts (and hence everyone bringing forth another) … and only then and because of this *such a product as an organised and self-organising being can be called a natural purpose*.
>
> In a watch one part is the instrument by which the movement of the others is effected, but one wheel is not the efficient cause of the production of the other. One part is certainly present for the sake of another, *but it does not owe its presence to the agency of that other*. For this reason, also, *the producing cause of the watch and its form is not contained in the*

[62] It is also worth bearing the binding problem in mind here to get a sense for how differently the "problem" is approached – it doesn't arise!

> *nature of this material, but lies outside the watch in a being that can act according to ideas of a whole which its causality makes possible*. Hence one wheel in the watch does not produce the other, and, still less, does one watch produce other watches, by utilizing, or organizing, foreign material; hence it does not of itself replace parts of which it has been deprived, nor, if these are absent in the original construction, does it make good the deficiency by the addition of new parts; nor does it, so to speak, repair its own defects. But these are all things we are justified in expecting from organized nature. – An organized being is, therefore, not a mere machine. For a machine has solely *motive power*, whereas an organized being possesses inherent *formative* power, and such, moreover, as it can impart to material devoid of it – material which it organizes. This, therefore, is a self-propagating formative power, which cannot be explained by the capacity of movement alone, that is to say, by mechanism." (Kant 2007: 202) (emphasis my own)

Following the ideas of cognitivism and AI xrisk, we might say that this "motive power" is "orthogonal" to the "formative" power. The motive power is the goal-independent power to execute the function, an exogenously specified function, that which specifies why it has the form that it does. The capacity to do the thing is orthogonal to why the thing matters. The watch "has solely motive power" and does not specify its own "end" or goal, and therefore why it has the form that it does. Much as the cognitivist account of cognition, the "watch" is assumed, and the relevant thing to explain is the functioning thereof. We can say more.

I want to draw attention in particular to the line "…it does not owe its presence to the agency of that other [part]." Here I am choosing to read Kant as identifying the determination of purpose of the watch as exogenous. It is not "forming" or organising itself according to its own "natural purpose". Any such purpose of the watch is exogenous. It does not define its function or form itself. By contrast, "natural purpose" shows up in living systems in the way in which an organism as a whole *exists for and by means of itself*. This is essentially the process of autopoiesis, to be discussed momentarily. Summarising the difference, "[a] lifeless thing does not metabolize; hence "its duration is mere remaining, not reaffirmation" (Jonas 1966: 81).

There are of course more ways than this to interpret the difference Kant is saying. I am not interested in insisting that this is necessarily the meaning he intended, only that it is possible to interpret it in the way I am suggesting.

Kant's thinking here offers leverage to establish a link. The link goes from the life-mind continuity thesis, which embeds and situates mind research in the context of biology, to a distinction between living biological systems and non-living systems in terms of the source of their goal-directedness. The goal-directedness in non-living systems is exogenous, and in living systems is endogenous. The first link from life-mind continuity to a differentiation of living and non-living therefore leads to autopoiesis, which explains this dynamic of self-directedness.

I mentioned that these passages turn up in several places, the connection being the curious recursion and self-reference at play. Stuart Kauffman's summary citation of the passage from his 1993 textbook *The Origins of Order: Self-Organization and Selection in Evolution* is worth quoting, in part because it makes for a nice summary, but also because he writes in the context of complexity science, a nascent field highly relevant to the discussion of autopoiesis to which I will now turn:

> "…the benchmark view of organisms in the Enlightenment was set by Kant, who undertook to distinguish organisms from mechanical devices. For Kant, organisms were fundamentally self-reproducing, and therefore *self-organising wholes*. In a mechanical device, the parts exist only *for* one another in that each is the condition of the other's functions toward a common functional end. In contrast, in an organism, the "parts" exist both for one another and *by means* of one another. For Kant, an organism "is that in which everything is both a means and an end". (Kauffman 1993: 4)

To clarify, when Kauffman says "self-reproducing, and therefore self-organising wholes", it could be maintained that these are not the same, or do not imply each other. There might be an intuition that a thing could create a new version of itself without being self-organising[63]. If self-reproduction is understood the way production of a vehicle is understood, or the way a robot building a duplicate robot would be understood, self-reproduction is not the same thing as self-organisation. That is, if the reproduction is of something *other,* then that is not strictly self-organisation, even if it is a robot building a duplicate. The two robots are not identical in the sense of being the same thing. They are two expressions of the same thing. In the subsection "Internal Norms of Self-Regulation" (§4.3.2.3.) of the coming section on Autopoiesis, this other-production is known as "*allo*poiesis". Compare this to an organism which, in self-reproducing, does not create a duplicate, but rather only itself. This is *auto*poiesis is self-production. The next section is devoted to making sense of this loop.

5.2.2. Autopoiesis

Now that the biological context established by life-mind continuity thesis is in place, the discussion of autopoiesis here will make more sense, particularly for bringing out the problem-defining character of post-cognitivism. Autopoiesis was first formulated as a notion by Maturana and Varela (1973, 1980, 1987) to address questions of organisation in biological systems. Maturana describes in (1973: xvii) in a personal manner the way in which the word came to him, more capable and potent in describing the dynamics of living systems than 'circular organisation'. Maturana puts it like this. Following a discussion with a friend on

> "…Quixote's dilemma of whether to follow the path of arms (*praxis*, action) or the path of letters (*poiesis*, creation, production) and his eventual choice of the path of *praxis* deferring an attempt at *poiesis*, I understood for the first time the power of the word 'poiesis' and invented the word that we needed: *autopoiesis*. This was a word without a history, a word that could directly mean what takes place in the dynamics of the autonomy proper to living systems…It simplified enormously the task of talking about the organization of the living without falling into the always gaping trap of not saying anything new because the language does not permit it." (Matura and Varela 1973: xviii):

[63] Thanks to my supervisor Elizabeth O'Neill for making this point.

It has since become integral to the biologically inspired models of cognition of Enactivism and, as we will see later, 4E cognition[64]. In Maturana's early work we see this move from a question of biological organisation to a question of cognition:

> "A cognitive system is a system whose organisation defines a domain of interactions in which it can act with relevance to the maintenance of itself, and the process of cognition is the actual (inductive) acting or behaving in this domain. Living systems are cognitive systems, and living as a process is a process of cognition." (Maturana 1970: 13)

It is this dynamic in which a cognitive system, by means of its organisation, defines for itself a domain of interaction in which it acts "with relevance to the maintenance of itself", that we see a strong departure from the cognitivist model. What we are beginning to talk about here is how an individual comes to be in an environment with a particular orientation in that environment based on its endogenous needs and capacities of perception. In this light we can jump straight in and note that autopoiesis is the emergence of a precarious identity, a process of self-creation in which a domain of significance for that identity is brought forth, co-emerging with the identity (Thompson 2007: 158). This is a more contemporary definition. In one of its earliest formulations, Maturan and Varela formalise it between summarising it as a process of homeostasis:

> "Autopoietic machines are homeostatic machines. Their peculiarity, however, does not lie in this but in the fundamental variable which they maintain constant. *An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production* [poiesis] *(transformation and destruction of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network*. It follows that an autopoietic machine continuously generates and specifies its own organization through its operation as a system of production of its own components and does this in an endless turnover of components under conditions of continuous perturbations and compensation of perturbations. Therefore, an autopoietic machine is an homeostatic…system which has its own organization…as the fundamental variable which it maintains constant." (Maturana and Varela 1973: 78-79)

Needless to say, there is much to unpack here. In the rest of the chapter, I bring out key features of contemporary understandings of autopoiesis which, while they have grown from this original presentation, have really kept the "fundamental variable" constant through time.

The first is a notion of self-producing wholeness. This is based on Kaufmann's reading of Kant on organisms, read as "fundamentally self-reproducing, and therefore self-organising wholes" (Kauffman 1993: 4). Kant's specific thoughts here appear and reappear across authors as Kant articulates with remarkable clarity the dynamic central to autopoiesis, namely, the way in which the whole regulates the

[64] Thompson's chapter 5 in *Mind in Life* charts in detail this trajectory up to Enactivism.

very production of itself. Now, this happens by means of, and according to, "internal norms of self-regulation", the second feature I will highlight (Thompson 2007: 152-153). The norms are internal in the sense that they are defined autonomously and endogenously, by and for the autopoietic individual. To illustrate this, I will pull from complexity science the notion of "constraint closure" (Montévil and Mossio 2015, Kauffman 2019), the third feature. Autopoietic systems are complex systems. I will fill out the discussion of constraint closure with more details of the thermodynamic "far-from-equilibrium" condition of autopoietic being. This sets up the last feature I wish to highlight, the way in which autopoiesis is a condition of "meaningful perspective". Autopoiesis is the emergence of a precarious identity to which things are endogenously meaningful to the individual to the extent that, and in the way in which, they make a difference to that individual, from their perspective. This means that, for organisms, the meaning of these differences is something *endogenously* determined, whilst for machines in the cognitivist model, the goal is already-given and externally specified. Things can be, and are, still meaning to the machine to the extent that, and in the way in which those things make a difference for the machine, *but the basis upon which that meaning is established is something exogenously determined.* As I speak of it here, "meaningful perspective" is *endogenously* determined.

"Meaningful perspective" thus builds on the way in which the organism autopoietically creates its own "constraints" and perspective. Understanding the difference between the meaning and perspective of cognitivist machines and post-cognitivist organisms therefore requires an appreciation for these things first.

If there is a thread through here, it is to do with the way in which autopoietic entities are a dynamic process of individuation. A whole individuates itself, according to norms of self-regulation which it itself determines. These dynamics produce *generative* constraints, generative insofar as they enable the reactions necessary to produce those very constraints. In this way the whole regulates its own production of itself.

What is relevant for my purposes here and worth highlighting in these features is that autopoietic entities have an *endogenous* sensibility for relevance and meaning. This is not something that computers or AI systems have because meaning for them is something *exogenously* determined. Neither computers nor AI are autopoietic.

Following the definition of "problem-defining", we can say that autopoietic entities enact a basis of orientation in the world, and on that basis are able to discern meaning of features they encounter for themselves. Such discernment is inherent to autopoietic beings. No less than the integrity and survival of the autopoietic individual depends on this capacity. In this way, and in terms of this thesis, a capacity to define problems is therefore fundamental as the very condition of being for enactivist and post-cognitivist conception of individuals. Their capacity for problem-defining is the dynamic of their very existence. Getting a sense for this is the aim of this section on autopoiesis. In order to do so, there are a few more angles we can take on autopoiesis. Because living systems are autopoietic and AI systems are not, the autopoietic mode of cognition is key to understanding much of the source of the differences between cognitivism and post-cognitivism. The following section on cognition as sensemaking makes much more sense if it is understood as a mode of cognition particular to this autopoietic condition of being, something computers and AI cannot do.

5.2.2.1. Self-Producing Wholes

Autopoiesis is a term that refers to a mode of biological organisation – self-creation (auto-poien) (Maturana and Varela 1998: 43). More precisely, it is the self-creation of a self that creates itself. It is a whole which produces itself. Echoing Kant above, autopoiesis identifies the dynamics of an individual which, as a whole, is generated by the way the individual, as a whole, regulates the interactions of its parts.

A commonly referenced example of autopoietic wholes is the manner of self-production of a biological cell "...a cell produces its own components, which in turn produce it, in an ongoing circular process" (Thompson 2007: 98; Froese and Ziemke 2009: 25). A basic approximation of the complexity of molecular dynamics in the autopoiesis of a cell holds that DNA codes for the production of proteins which build a cell wall (Thompson 2007: 97-102). This cell well operates as a constraint, enclosing a domain and constraining to within the boundaries of the cell those biochemical reactions necessary for the very existence of the cell. A cell does - *is* - by virtue of it defining its own constraints. It defines those constraints which enable the condition of itself. Popularly, "life creates the conditions for life".

Thompson formalises these circular dynamics with three criteria:

> "For a system to be autopoietic, (i) the system must have a semipermeable boundary; (ii) the boundary must be produced by a network of reactions that takes place within the boundary; and (iii) the network of reactions must include reactions that regenerate the components of the system.
>
> In summary, the form or pattern of the autopoietic organization is that of a peculiar circular interdependency between an interconnected web of self-regenerating processes and the self-production of a boundary, such that the whole system persists in continuous self-production as a spatially distinct individual." (Thompson 2007: 101)

The circularity, recursivity, and self-reference stands as an interesting provocation of intuitions about linear causation: a self emerging from conditions created by itself; a self-emerging self emerging. Escher's *Drawing Hands* (fig. 4.1.) below expresses the same defiance in the way that the very conditions necessary for there to be any hands drawing at all are brought about by those very hands. There is no "in the first place". This is perhaps a helpful way of understanding just what "enactive" of the Enactive theory of mind means – in producing the conditions which produce itself, an individual enacts itself and so brings itself into being. Recall in the chapter on cognitivism the idea that showed up in Gödel, Turing's Universal Turing machines and the Halting problem, Russell and Whitehead's *Principia Mathematica* – strange things happen when things take themselves as objects of their own consideration, in their own way. The enactive account does not try to square the circle, but puts it front and centre.

Now, this commentary may offer more confusion than clarity. It can be taken as an expression of a *style* of the kind of thinking of the paradigm. As far as this thesis is concerned with the ways of thinking available to us, this thesis mostly focuses on *what* the two different paradigms think. If nothing else, this commentary points to a bending and curving of the straight lines and orthogonal angles of cognitivism. A circling of the square. It is more than just an aesthetic difference, but that may be a sufficient way to describe it for now.

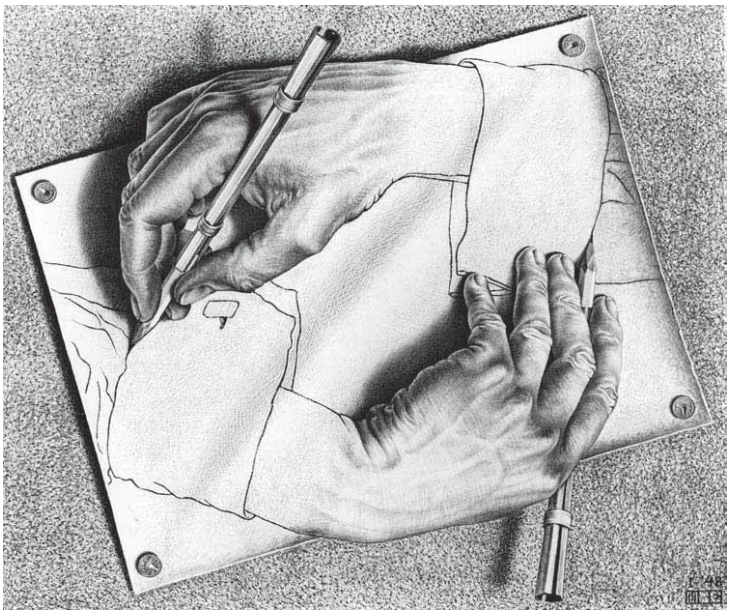


Figure 7. Drawing Hands, M. C. Escher, January 1948, Lithograph
https://mcescher.com/gallery/back-in-holland/#iLightbox[gallery_image_1]/32

Returning to the discussion, the semi-permeable membrane of the cell is significant for the way in which it both individuates the individual and operates as a means of continuously maintaining the integrity of the individual by differentiating and defining what is allowed across the threshold from the environment. Because it involves discerning and defining what is to be metabolised and what is to be blocked, the existence of a semi-permeable membrane speaks directly to problem-defining as a fundamental characteristic of autopoietic systems. The membrane is *semi-permeable*, meaning it is open enough to allow passage across the boundary of nutrients and waste, and closed with enough integrity to distinguish it as an individual from the environment in which it is embedded.

> "The membrane serves as a barrier to free diffusion between the cell and the environment, but also permits the exchange of matter and energy across the boundary. Within the boundary, the cell comprises a metabolic network." (Thompson 2007: 98)

The membrane is the boundary drawn, generated by the cell, which individuates it as such – "…without the boundary containment provided by the membrane, the chemical network would be dispersed and drowned in the surrounding medium...a cell stands out of a molecular soup by creating the boundaries that set it apart from what it is not" (ibid: 99). In this way the semi-permeable membrane is the figurative line, surface, and threshold which separates and regulates biochemical commerce between a self-produced distinction between an "inside" and "outside". More to the point, the production of the boundary and the regulation that occurs at its threshold speaks, again, to the "discernment of meaning"

of problem-defining because the regulation that occurs at the boundary requires a capacity to discern, from the perspective of the cell, the "meaning" for itself of the molecules encountered, which is to say, the particular significance they take on for the cell; whether they should be let in or blocked.

A final point to note is a distinction between first-order and second-order autopoietic systems: "Living cells are first-order autopoietic systems, whereas systems that include individual cells as structural components are second-order autopoietic systems." (Thompson 2007: 105). Thompson then suggests that the interesting question is "whether any second-order autopoietic systems are also first-order autopoietic systems" (ibid). I will not go into this other than to note that, in reference to his own criteria which I noted above, what matters is a certain organisational closure, one effected by a boundary, membrane or otherwise. A second-order autopoietic system is also a first-order autopoietic system if it expresses the same Kantian holism I discussed above.

Understanding autopoiesis as the dynamic of a self-producing whole avoids some of the problems cognitivism faces, in particular, the binding problem. Saying a bit more about this offers a means of further differentiating post-cognitivism from cognitivism as well as clarifying the significance of "self-producing wholes".

5.2.2.2. Binding Wholes

The point of interest here is to do with the way cognitivism has to assume as already-given the very agent in order for its model of cognition to get going. Insofar as cognitivism assumes this, it does not account for it, and so is left with the now familiar sense of arbitrariness expressed is the binding problem. In contrast, by including the dynamics of the individuation of the agent, the autopoietic production of both parts and wholes, binding is not a problem for autopoietic accounts of cognition. For the purposes of this thesis, this point serves only to bring the characteristics of both paradigms into greater relief with respect to one another. Autopoietic systems of post-cognitivism define themselves, part and whole. Focusing primarily on how cognition solves problems, cognitivism takes as already-given the definition of the agent in order to get off the ground. It has to assume something outside the system defines the system. At a risk of overlabouring the point, this means to say that where defining occurs endogenously in autopoietic systems, it occurs exogenously in cognitivist systems.

This is the move where the arbitrariness comes in. Because the only way to account for the individuation and identity of the problem-solving agent and the problem is exogenously by the modeller, there is no way to account, for the endogenous *significance* of the identity of the agent, or the function it is executing. Thompson notes that in this way, the binding problem is not a problem of the individual in question, but of the observer who happens to be theorising about the individual (Thompson 2007: 53). This in turn prompts questions about where the limits of "the system" in question really lie[65].

[65] See (Barad 2007) for an in-depth book-length work on this, at the intersection of feminist philosophy of science and philosophy of physics. The philosophy of science question is closer to the philosophy (and ethics) of AI question than might appear at first glance and has implications for a philosophy of technology too. If nothing else, it is revealed that these three fields are connected in ways that deserve further exploration. The philosophy of science question, explored by Barad, concerns the way in which factors ostensibly "outside" the experimental setup are necessary conditions for the particular observations of the experiment. (See footnote 20 on the Stern-Gerlach experiment on page 112.) The observer is shown to be a necessary feature of the experimental set up. The analogue in the philosophy of AI is the way in which, in the cognitivist paradigm, the engineer building the system is taken to similarly be "outside" the system, and yet this engineer is the one who specifies the task (the frame),

Consider again Kant's watch example in terms of Kauffman's discussion of "Kantian wholes". In Kant's watch, the ordering of the parts, the integrity of the whole, and the particular way in which the parts are bound together are exogenously defined. "For this reason, also, *the producing cause of the watch and its form is not contained in the nature of this material, but lies outside the watch in a being that can act according to ideas of a whole which its causality makes possible*." (Kant 2007: 202) (emphasis my own). The whole according to which the parts are bound "lies outside the watch in a being" whose idea thereof is the source of the wholeness.

Where autopoiesis is concerned then, speaking of wholes therefore offers a means of distinguishing cognitivism and post-cognitivism in an important way. Insofar as cognitivism has to assume these things and make arbitrary definitions of wholes, we get both the interesting ideas like the space of possible minds, but also the binding problem. "Feature" and "bug". Insofar as autopoiesis is characteristic of post-cognitivism, it accounts for how both parts and wholes show up together and so does not suffer the same sense of arbitrariness. Kantian wholes show up bound together from the moment they spontaneously emerge (Kauffman 2019). By way of summarising the distinction, the problem cognitivism has which autopoiesis accounts for in post-cognitivism, is the famous grin of the Cheshire cat in *Alice's Adventures in Wonderland*.

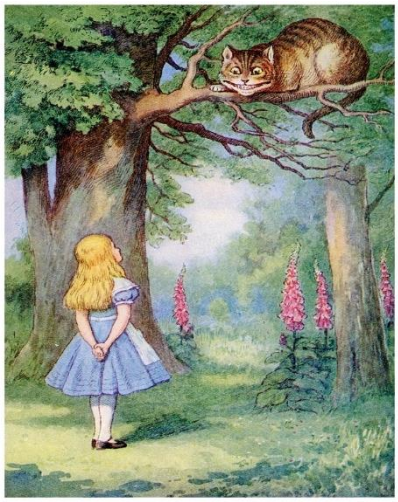


Figure 8. Alice and the Cheshire cat in Alice's Adventures in Wonderland

> As their dialogue comes to a close, part by part, the Cheshire cat disappears until all that remains is the famous, unembodied grin. (Carroll 2009: 75)

The Cheshire cat's grin is at once a part of the Cheshire cat, and yet cannot exist without the whole. A smile doesn't exist without a face, nor a face without a body, nor a body without an ecosystem, nor an ecosystem without many, and so on. Hearts, lungs, organs, "pumping", "breathing" are all functions which do not exist independently of the whole. In this regard, the binding problem for cognitivism

designs and "Binds" together the elements required for a machine to execute the task. The engineer specifies the "why" and yet is excluded from the model of the system pursuing "its" goal. This leaves us then with a model of cognition in which Orthogonality and Instrumental Intelligence make sense and the proposition of optimisation of the task leads to the notion of Superintelligence. When all these are put together, then we have xrisk.
The analogue between philosophy of science and philosophy of AI here then is that, in the case of the scientific method, the otherwise excluded "observer" sets up a condition of observation and awareness, and in the case of AI, the otherwise excluded engineer sets up a condition of agency, and in both cases, they are necessary to explain the significance and relevance of the results – the "why". The connection to philosophy of technology is that this exclusion influences our practice of characterising AI. When we ignore it, it makes it easier to grant agency. When we recognise that the human is a necessary feature in the model, it is easier to see AI as a tool and not an agent. This intersection deserves more exploration than this footnote.

reveals a conception of things that is much like the Mr. Potato Head toy in which the modular facial features – eyes, ears, mouth, and nose and so on – can be plugged into Mr. Potato Head's body in any entertaining arrangement. The parts exist independently of the whole and can, even in just in the space of possible Potato Heads, be arranged in more or less any fashion. Echoing the Orthogonality Thesis (Bostrom 2012, 2014) in which the value of modular functional parts of cognition like "sentience" and "sapience" can vary freely of one another, the binding of eyes and ears etc. is not a necessary coupling. They are defined parts whose values can vary independently of one another. Repeating from the last chapter, from the perspective of post-cognitivism, this makes about as much sense as having leaves, flowers, stems, and branches, and trunk and roots, and saying that in the "space of possible tree functions", more or less any combination of these parts is in principle possible.

To summarise, in this first section on autopoiesis I sought to define autopoiesis in terms of wholes which produce themselves. The discussion was centred around a reading of Kantian wholes which *exist for and by means of themselves*. This is a reading which shows up in many places, including those which specifically engage with autopoiesis in the context of the enactive theory of mind (Thompson 2007, Froese and Ziemke 2009) as well as those which explore it in the context of self-organisation, complexity and evolutionary theory (Kauffman 1993, 2000, 2019). This framing of autopoiesis as wholes produces an important distinction between cognitivist and post-cognitivist ways of thinking. Understood in terms of wholes, autopoiesis begins to account for the individual that cognitivism takes as already-given (and then subsequently is forced to face as an externality in the guise of the binding problem). Cognitivism notes there is no in-principle reason why particular parts need show up together and then has problems explaining why certain parts do in-fact show up together as a wholes. I claimed that this is because autopoietic wholes are self-producing, producing the parts and wholes in reciprocal fashion.

As a reminder, the notion of autopoiesis is important for the purposes of this chapter for the way it expresses and supports a characterisation of post-cognitivism in terms of problem-defining. Framing autopoiesis in terms of self-producing, "Kantian wholes" is a first step in making this clear.

This framing can now be further complimented with a framing from complexity science, since autopoietic systems are also complex systems. There are several things a complexity framing of autopoiesis brings to the discussion. A specific feature I want to bring in is that of "closure". I mentioned above that autopoiesis involves "organisational closure" (Thompson 2007). In the next section this can be filled out in more technical detail with the notion of "constraint closure", thermodynamic work cycles and "work closure". What is specifically significant about closure is the way in which it involves a self-determination and definition of an individual. The rest of the discussion from complexity is important for the general picture it creates of the necessary precarity of autopoietic individuals. This precarity is the condition of existence of autopoietic entities and, anticipating the rest of this chapter, speaks to how autopoietic individuals make sense and meaning of the world. As a reminder, these dynamics of individuation of the agent are all things which the cognitivist model takes as already-given. From the perspective of enactivism, the cognitivist model of cognition begins only once this process of individuation has generated an agent, but, as we will see in throughout this chapter, why and how the agent comes to encounter meaning in the world from their perspective, is entirely to do with this process of enactive individuation that the cognitivist model excludes.

5.2.2.3. Complexity and Constraint Closure

There are several points to the trajectory of this section, each of which build on each other, eventually producing a picture of the "constraint closure" enacted by autopoietic systems. In autopoietic fashion, constraint closure conducts a thermodynamic work cycle in which the constraints necessary for "work" to happen are produced by the very "work" they afford. Closure speaks to the (now) thematic wholeness of systems which exist "for and by means of themselves". This is to say, understood in terms of complex systems, self-defining comes out again as integral to the autopoietic process with which post-cognitivism begins to make sense of mind and cognition. The biologists Montévil and Mossio who introduced the notion of constraint closure suggest that whilst autopoiesis is vital to an account of this "self-determination", it does not fully account for the organisational closure fundamental to it. I have quoted them in full to begin this section:

> The centrality of organisational closure and its connection to organisation, as well as its distinction from (and complementarity to) thermodynamic openness, have become givens in most subsequent accounts of biological self-determination (Letelier et al. 2011). One of the best-known formulations is the one centred on the concept of autopoiesis (Varela et al., 1974; Varela, 1979) which, among other aspects, emphasises on the generative dimension of closure: biological systems self-determine in the specific sense that they "make themselves" (auto-poien). Precisely because of their dissipative nature, the components of biological systems are maintained only insofar as they maintain and stabilise not just some internal states or trajectories, but the autopoietic system itself, as an organised unity…
>
> In spite of its qualities, however, the concept of autopoiesis… suffers in our view from a central weakness, insofar as it does not provide a sufficiently explicit characterisation of *closure*. Biological systems are at the same time both thermodynamically open and organisationally closed, but no details are given regarding how the two dimensions are interrelated…" (Montévil and Mossio 2015: 2) (emphasis mine)

A helpful point of departure here is the notion of the semi-permeability of the autopoietic membrane. In the language of complexity, such a system is spoken of as being *thermodynamically open* (Kauffman 2019, Krakauer 2019, Mitchell 2009). This just means that system is not sealed, and that instead there is passage across its boundary of energy, matter, and information. From the perspective of the system, it lets things in and out. In his discussion of autopoiesis, Evan Thompson (2007), also citing Francisco Varela (1979: 55-60), speaks of organisational and operational closure in thermodynamically far-from-equilibrium systems, so there is a tight conceptual knit of the concepts here between the thermodynamics, complexity science, and biology of living systems (Thompson 2007: 45).

There are a few basic things to say about thermodynamics. It was discussed in the previous chapter in the context of Shannon's information theory because information and entropy are sides of a coin. A quick introduction again will ensure a generative context for what is to come. Thermodynamics is a branch of science concerned with how the energy of a system changes, understood in terms of concepts of like heat, temperature, entropy, and work. A system at a state of rest is said to be in equilibrium. This

means there is symmetry between it and its environment – across the boundary – like when an air mattress or balloon deflates until the pressure inside the mattress is equal to the pressure outside. A system that is transforming is at a non-equilibrium and so there is a "symmetry-breaking", a difference between the sides of a boundary (Krakauer et al. 2020). Information is often understood in this way, as a place of broken-symmetry and so catches our attention, like the "third man in the bar" joke.

Information flow is interesting in this context too because whilst some systems move towards equilibrium and symmetry, living, autopoietic, complex systems maintain themselves in a necessary non-equilibrium, maintaining a broken symmetry. Living systems are not in thermodynamic equilibrium with their environments in the same way that an inflated air mattress is not. This means that living systems are highly information-dense. Their boundaries mark the edge of great difference. The difference is that a mattress is an exogenously individuated artefact which does not produce itself the boundaries which hold it at in non-equilibrium. Much as autopoietic systems do, complex systems produce their own boundaries which enable them to hold themselves at non-equilibrium. They are self-determining (Montévil and Mossio 2015). I will continue to make this point throughout this section and chapter. Note though, whilst all autopoietic systems are complex, not all complex systems are autopoietic. Hurricanes and tornadoes and "dissipative systems" which do not maintain themselves, are not autopoietic.

At this point I want to note that in all this discussion a boundary is assumed. It is the sides – the "inside" and "outside" of an individual's boundary – between which there is a not-symmetry. This is what enables us to speak in terms of countable individuals or determinable "systems" rather than open "fields" of dynamics. Such fields tend to follow the second law of thermodynamics and be analysed in terms of how they reach a stable equilibrium of maximum entropy in their respective media, e.g.: air pressure and fluid dynamics. By contrast, the thermodynamically open systems in question are not equilibrium systems. Referencing Nobel laureate chemist Ilya Prigogine (Prigogine and Nicolis 1977), Kaufmann notes that thermodynamically open systems are "displaced from equilibrium" and that:

> "…such systems can "eat" the order in their environment, such as gradients, and build order. Non-living systems like whirlpools and Bernard Cells, in which convective flow patterns emerge spontaneously in plates of viscous fluid slowly heated from below, show that *pattern can emerge in such systems while displaced from equilibrium*." (Kauffman 2019: 18-19, my emphasis)

The difference here, between non-living, "dissipative systems" and living – autopoietic – systems, is "propagation of organisation" (Kauffman 2019: 19). Both living and non-living systems here are complex systems, but the organisation of these non-living systems dissipates as they release energy into the environment according to principles of entropy. By contrast, living systems constrain energy within their boundaries, and in a way that produces the very constraints to do so. This enacts a propagation of organisation.

That such complex systems exist *and maintain themselves* at non-equilibrium is now a basic premise in understandings of biological organisation and complexity (Montévil and Mossio 2015), but for the purposes of this thesis it is non-trivial. The goal of this chapter is to legitimise a characterisation of the post-cognitivist paradigm in terms of "problem-defining". Autopoietic systems are effectively self-defining systems. What complexity brings to the table is to show that, as complex systems, autopoietic

systems not only define themselves, but thermodynamically hold themselves at the non-equilibrium necessary to do so.

Autopoietic systems like a cell are therefore thermodynamically "far-from-equilibrium"[66], emerging and operating in a very narrow space on the thermodynamic spectrum between order and chaos at what is called "criticality" (Bak and Paczuski 1995, Mora and Bialek 2011). This archetypically "Goldilocks" space is the precarious middle in biology, chemistry, and physics between the rigidity of order and the incoherence of chaos. It is precarious because it is not an equilibrium state, not a state of rest, but a state of transition and change. It is a thermodynamic state that does not remain unless work "holds" it there, so to say. The poles of the thermodynamic spectrum, order and chaos, are equilibrium states. This means that, where the instability of criticality means a high possibility of change and transformation, a condition required for autopoiesis, in their respective ways, the equilibrium states of chaos and order mean that chemical reactions and transformation are unlikely. The precarity of criticality is necessary for autopoiesis.

A typical example of a fully ordered state is that exhibited by a crystal or diamond. Their structure is ordered, and so rigidly-so that reaction and transformation are all but impossible, requiring huge amounts of energy to disrupt the constraints which preclude reaction. A typical example of a chaotic state is one exhibited by uncontained gases. There is no structure constraining the degrees of freedom of the gas. It is only the bonds between the particles and molecules holding it together, and so the gas dissipates into its environment. We can say therefore that the self-determination and existence of autopoietic, complex systems relies on a capacity to hold themselves in a precarious (far-from-equilibrium) dynamic…so that they are able to continue to do so. I find it powerful to imagine an object like a glass at equilibrium resting on a table and, at the other end of things, also at equilibrium falling through the air, and then in between, at meta-stable equilibrium precariously wobbling on the edge of the table, somehow holding itself there, poised in possibility between lounging on the table and smashing on the floor. This is not a stable dynamic, but it is "meta-stable" insofar as this precarious instability is something that is stably maintained:

> "…biological systems can be understood in terms of "extended critical transitions", which mean that they form coherent structures, whose proper symmetries are inherently unstable. … In contrast to the role played by theoretical symmetries in the mathematical and theoretical definition of physical objects, their instability in the biological domain underlies the fundamental contextuality, variability and historical nature of biological phenomena." (Montévil and Mossio 2015: 6)

To understand how this propagation of organisation at precarious far-from-equilibrium happens, a basic understanding of the notion of "work" in the physicist's sense of the term is helpful.

Kauffman talks about work as "force acting through a distance" and uses the example of a hockey stick moving a puck (2019: 19). Here work is energy expressed as the force moving the puck "through a distance". It is the expression of energy through particular degrees of freedom, the particular directions in which the puck is moved (Atkins 1984). Whilst these constraints may not be the exciting part of the situation, being just the walls off which the ball bounces or the stick off which the puck moves, work doesn't in fact happen without these constrained degrees of freedom. Kauffman writes:

[66] See (Ornes 2017) for an overview of the concept and its significance across the sciences.

> "Think of a cylinder and piston, with the "working gas" confined to the space between the cylinder and piston. The expanding gas does work on the piston to move it in the cylinder. This is the constrained release of energy into a few degrees of freedom.
>
> For a physicist, "degrees of freedom" mean roughly what is now possible. In the absence of the cylinder, the hot gas would expand in all directions. No work would be done. In its presence, the gas expands only along the cylinder, thrusting the piston. And so work is done." (Kauffman 2019: 20)

Now the interesting question is who or what defines and constrains those degrees of freedom. This is not included in a notion of work, in the same way that cognitivism conceives of cognition. Work talks of quantities like mass, acceleration, and force, but it says nothing about why I might move the puck in the particular direction I do, which is to say, why it is those particular degrees of freedom through which energy is expressed. *The notion of work assumes the constraints*, it implies constraints, *it takes them as already-given.* This is not intended as a critique of the notion, only to show that it makes the same move as the cognitivist model of cognition.

The comparison with cognitivism here is, I think, worth exploring. The already-givenness of the constraints necessary for the process to occurs resembles the already-givenness of the cognitive situation. A comparison with the Orthogonality Thesis is also possible. We can say that energy and the particular degrees of freedom through which it is expressed as work are variables which can vary independently of one another. The same sense of arbitrariness is there. We might say "there is no reason why energy and particular degrees of freedom necessarily go together in the space of possible work."

In this regard, the constraints necessary for work amount, if functionally, to an account of "why" the work was directed in the particular direction that it was. This point should not be overthought. I don't want to suggest here that constraints amount to an epistemological explanation of the work or anything like that, even if that argument can be made. The significance of the constraints here is simply that they bring out the main distinction between the cognitivist and now post-cognitivist models I have been describing. In the cognitivist model of cognition, the claim is that the "cognitive situation" is taken as already-given. In the cognitivist model, the cognitive situation is, in effect, the set of constraints in which cognition is realised as the particular perspective of an agent in a world they perceive. In articulating these constraints in our cognitivist model of such an agent, we are defining degrees of freedom of possible action or cognition, specific possible expressions of (cognitive) work. In much the same way that work presupposes constraints, cognition presupposes what I have been calling the "cognitive situation". This is to say, then, that a model of work is much like a model of cognition. Much as we cannot account in a model of work for why the work is directed the way it is without reference to the constraints, we cannot account in our model for why the (cognitive) work is directed in the particular direction that it is without reference to the "cognitive situation". It is in this sense that I speak of the "why" of the work and the way in which it is independent of the force of work itself."

Now, one thing we can now notice is that, whilst these things are conceptually distinguishable, they are not in fact[67]orthogonal to one another because the constraints are absolutely necessary for the possibility of work. Without the constraints of the "why", there is nothing to channel the energy, so it dissipates.

As Kauffman puts it:

> "[t]he physicist cheats when he or she just puts in boundary conditions [i.e. the constraints] for the cylinder and piston and leaves it at that. After all, just where did the cylinder come from since the Big Bang? Well, it took work to construct the cylinder that then serves as a constraint on the release of energy. It took work to construct the piston. It took work to assemble the piston inside the cylinder and arrange for the gas to be at the head of the cylinder. Physicists ignore this when they merely impose boundary conditions with no consideration of from whence they came." (Kauffman 2019: 21)

Note again the sense of arbitrariness which Kauffman exposes. From the perspective of enactive autopoiesis, cognitivism seems to be doing much the same thing here, "imposing boundary conditions with no consideration of from whence they came". The already-given cognitive situation are the particular constraints through which the intelligence, energy, and agency of the agent are expressed as work to solve the problem in question. But it took work to produce those constraints in the first place, work which cognitivism excludes from its model.

I want to emphasise this point because it maps onto the core cognitivist assumption in an interesting way. The point then is that work does not account for the means of its own possibility or production. It requires already-defined constraints. Cognitivist cognition too requires already-defined constraints – the cognitive situation in its case. Put in the terms of work and constraints, the already-defined agent of cognitivism can be thought of as a complex of constraints on the degrees of freedom of their agency, choice, energy and so on, whilst their problem from their perspective can be thought of the direction in which they choose to direct their individual agency. Meanwhile, the agent is embedded in a context – a situation – which also places constraints on the degrees of freedom of the agent's possible action. Putting this all together, the analogy or mapping is as follows. Cognitivist cognition describes a process of work given constraints[68]. What autopoiesis offers, understood from the perspective of complexity dynamics here, is an account of the generation of those constraints.

Work and constraint-closure have I think, much to offer for making sense of autopoiesis and the way it leads to a different model of cognition than what we get with cognitivism. I want to bring this out as we continue the thread of this section.

The value of autopoiesis here then is how it accounts for the production of the very constraints which produce it. In the context of complexity, this is done by building on the notion of "work" to the notion

[67] In an earlier footnote, (50), I quoted Daniel Dennett's council that "One should be leery of these possibilities in principle…sometimes an impossibility in fact is theoretically more interesting possibility in principle..." (Dennett 1991: 4).

[68] Whilst I do not want to go into any more depth here with this analogy, I do think there is a lot of substance worthy of further consideration outside this thesis. In a paper being prepared for journal submission, I discuss in more detail a mapping between cognitivism's problem-solving conception of intelligence and thermodynamic work, and a possible post-cognitivist conception of intelligence as mapping on to thermodynamic "work-cycles" and "work-closure", concepts I turn now to discussing here.

of a "thermodynamic work cycle", or, as Kaufmann (2019: 21-29) refers to it with minor differences, a "constraint work cycle". In autopoietic fashion, work is constrained back into the construction of those very constraints, which then go on to afford the work which builds those very constraints. When this loop closes back on itself, the closure effects a cycle of work in which order is propagated:

> "…a set of constraints on a set of non-equilibrium processes can achieve a *work task closure* that constructs the very same set of constraints. The constraints do work tasks that construct the same constraints or boundary conditions.
>
> The system can literally build itself! This is the amazing concept of Constraint Closure of Montévil and Mossio (Montévil and Mossio 2015)." (Kaufmann 2019: 22)

This loop closure of the work cycle is the difference between those dissipative structures mentioned earlier in which, without constraints enclosing energy, they lose energy to the environment, dissipating as they do so. It is the way systems build themselves up in complexity, locally "beating" the second law of thermodynamics. Strictly speaking the second law is still at work universally, but is put to work in a way that is generative of further work and complexity, generative of possibility rather than being an expression of terminating possibility:

> When a work cycle (W-C) is realised, constraints which apply to the system are not independently given … but rather are produced and maintained by the system itself. Hence, the system needs to use the work generated by the constraints in order to generate those very constraints, by establishing a mutual relationship, i.e. a cycle, between constraints and work. (Montévil and Mossio 2015: 4)

To summarise and to conclude the discussion about constraint closure, we can consider it as the difference between a fire in your hearth, and your home being on fire[69]. The hearth operates as a condition of constraint on the degrees of freedom of the fire. The barrel of a gun or a cannon does precisely this too, defining and constraining the degrees of freedom through which energy, in the form of the projectile, is directed. Engines, too, house a reaction which produces energy which, when constrained through appropriate degrees of freedom propel vehicles and machinery. The difference between heating systems, projectile weapons, and engines, on the one hand, and organisms, on the other hand, is the same difference to which Kant, Kauffman, and Montévil and Mossio speak. There is in these technological, machine systems no self-production and no self-generation. The organisation of the artificial system is not self-determined. A rifle does not define its own constraints, nor does an engine, or a fireplace, or any other such non-living tools. A cell does. Complex systems do. Autopoiesis does.

Connecting this to cognition, we can say the following. All machines perform work, but not all machines effect a work *cycle* as autopoietic, living systems do. If there is a fundamental distinction between machine systems and living systems, it seems like this might be the place. (There is nothing in the concept of "machine" which analytically prohibits the conceptual possibility of a machine effecting a work cycle, but for now "machine", by contrast to "living system", captures a system that requires

[69] I am quite sure the credit for this example is owed to Stuart Kauffman, but I can no longer trace it.

exogenous work for its possibility.) Computation describes the "work" which we call cognition, but not the cycle. It is not a complete "cognitive work cycle", if you will. Autopoiesis is the dynamic of the complete cognitive work cycle.

This section on complexity and constraint-closure and work-closure served to reiterate the point that autopoietic systems define themselves endogenously and machine systems like AI, and the technological objects with which cognitivist thought has heretofore been concerned, are exogenously defined. I think this difference between endogenously and exogenously defined individuals matters. In this way I hope to begin to make sense of the characterisation of the post-cognitivist paradigm as problem-defining.

We can now move from the question of how this organisation happens, to what it means for the individual. That all this organisation is internally regulated is important. It means that, from the perspective of the self-maintaining individual, certain things are *endogenously* more and less relevant. This means that certain things matter to it, but that it must discern just in what way, and to what extent such things are of value to it. Given a notion of autopoiesis, complimented with notions of constrained degrees of freedom, a consideration of these endogenous norms of regulation begins to foreshadow the discussion of "sensemaking", the next concept of this chapter after autopoiesis.

5.2.2.4. Internal Norms of Self-Regulation

This chapter has gotten a lot of mileage so far out of the distinction between endogeneity and exogeneity. In this section, I want to continue exploring the significance of this difference, this time in terms of the "internal norms of self-regulation" according to which autopoietic systems produce and maintain themselves. Recall that autopoiesis is the emergence of a precarious identity, a process of self-creation in which a domain of significance for that identity is brought forth, co-emerging with the identity. What the autopoietic individual encounters therefore, is what is endogenously significant to it. This is in contrast to machines and existing AI systems for which the significance of what they encounter is exogenously defined for them. Here the significance is explained in terms of the difference between autonomy and heteronomy, and autopoiesis and allopoiesis.

By way of easing into this section, here is a quick orientation of where we are in this chapter on postcognitivism. In the context of the discussion about autopoiesis, these "internal norms" facilitate a move from talking about an autonomous condition of being, that is, autopoiesis, to how that condition of being is one which necessarily involves a sensitivity for meaning and relevance for the agent in question and *from their perspective*. From there it becomes possible to talk about how autopoiesis, and the other concepts I use to characterise post-cognitivism, invite an understanding of post-cognitivism in terms of problem-defining.

In the previous section I spoke a lot about the notion of closure. In this section the notion of autonomy will be key and putting it together with the notion of closure means we can talk about autopoiesis in terms of "internal norms of self-regulation" as a *closure of autonomy*. Here the conceptual work "closure" does is to distinguish between those technologies we call autonomous, and which conduct themselves in some sense, but are not autonomous insofar as the norms according to which they conduct themselves are exogenous. The norms according to which they regulate themselves are external. Recall Kant's line I quoted in the section on autopoietic individuals as "self-producing wholes" – the wholes exist "for and *by means of* themselves". Technologies we might consider autonomous insofar as they

move around a room and vacuum the dust autonomously, for example, lack this second aspect of autonomy in which they *exist by means of themselves*. By contrast, autopoietic systems have endogenous, "*internal* norms of self-regulation". Where the behaviour of the robot is autonomous once it is turned on, the norms according to which it regulates its behaviour are exogenously determined by the engineers who built it. Autopoietic systems endogenously regulate their own behaviour according to, again, endogenous norms.

"Internal norms of self-regulation" is thus another voice echoing what has been said in the previous sections on "self-producing wholes" and "complexity and constraint closure", namely, that the endogeneity of an autopoietic system is significant. The cognitive systems of cognitivism, at least as far as they are modelled, do not have this feature. Because their individuation is taken as already-given, the source of the norms which produce them in the first place and then regulate their behaviour, are excluded from the model.

To be a bit more concrete about what "internal norms of self-regulation" means, I want to reconsider the dynamics in terms of Kauffman's thermodynamic "work cycle" I mentioned in the previous section and used as an example the autonomy of a Roomba robot (Kauffman 2019: 21-30)[70].

The example Kauffman uses to set up the notion of a "work cycle" is a cannon firing a cannonball. The form of the cannon provides the constraints through which work is expressed, initiated by the powder explosion, projecting the cannonball through the air and then eventually into the ground, where it makes a crater with the left-over energy. There is no work cycle. In order to fire the cannonball again, it must be returned to the cannon, and of course the production of the cannon took work, which is not accounted for. The autonomy of the Roomba robot, or Turing machine, is analogous to the cannonball, more sophisticated in detail, but the same in kind. The constraints of the "cannon" through which the "work" is expressed, "firing" it out into the world whereupon it "flies" autonomously, are exogenous, as are the cannonball's.

Now, obviously I think we would want to say the robot is more sophisticated than the cannonball when it (the robot) is "in flight". The relevant point of analogy is the way in which there is no closure in both cases. The "work" which both cannonball and robot express dissipates into whatever task they are executing, rather than into furthering their own possibility of themselves, and so both must be "plugged" back into an external energy source, (a human, in the case of the cannonball), in order to be able to "function" again. The "work cycle" is completed by a human. There is not "closure". Now, organisms also get their energy from things outside their boundaries, but rather than being exogenously plugged in, they enact their own closure autonomously. They are responsible for getting their own energy, and their capacity to do so. There is "closure" to the "work cycle" of an organism. So, where the energy comes from is not the significant part of "exogenous". In the section on computation in the previous chapter, I suggested that computation was a description of a smaller arc of a larger circle – the "work cycle".

There is no closure of the work cycle with the robot and the cannon, or, more generally, for a Turing machine. Something external has to complete the circle. The tasks express work, but in an extractive way because they do not contribute to the regeneration of the possibility of work. As Kant points out in the case of the watch, whatever closure into a "whole" exists here "lies outside the watch in a being that can act according to ideas of a whole which its causality makes possible" (Kant 2007: 202). This means

[70] Everything which follows can be applied to Turing Machines too, at least, such as I presented them in the chapter on cognitivism, which is to say, what I go on to say in theory applies to cognitivist systems generally.

that "whole" is realised exogenously, in the minds of humans who find it relevant to build such things, and the extractive chains of production which produce them.

That such things do not define themselves leads to two important things in the context of this section. The first is that the frame of reference of the robot or cannon or agent is exogenous to it, because it is "by means of another" external agent that what is meaningful for the agent is defined.

Contemporary AI works like this. Recall from the discussion of AI existential risk what Stuart Russell says: "[t]he way we build intelligent agents depends on the nature of the problem we face." (Russell 2019: 43). This means that the individuation of agent is exogenously defined according to the problem it is intended to solve, something also exogenously defined, and all relative to a frame of reference of the engineering observer. Russell speaks to this too: "Because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build optimizing machines, we feed objectives into them, and off they go." (ibid: 10).

Harking back to the first chapter of this thesis, the frame of reference of Russell's thinking is "human compatible AI", AI that is "Safe", that is not a risk to humanity. In this frame of reference, what is relevant is not only building AI to human-level and beyond, whatever that looks like, but doing so in a way that is compatible with the flourishing of humans and the rest of the planet. It is in this context that I wish to quote Russell. Noting for starters that "humans are intelligent to the extent that our actions can be expected to achieve our objectives" (Russell 2019: 9), he then goes on to suggest that we don't want machines that are intelligent in the sense that we are. We would benefit instead from defining AI in terms of its capacities to achieve *our* objectives. He suggests we conceive of machines as "beneficial to the extent that their actions can be expected to achieve our objectives." (ibid: 11).

The objectives are externally defined, even if internally represented. The frame of reference in which those objectives show up as relevant to the machine is not the frame of reference endogenous to the machine, but the perspective that is externally projected onto it by the person who built the object. What is "meaningful to the machine" is exogenously given to the machine. What the watch, cannon, or Turing machine are doing is not endogenously relevant to them. It requires an external frame of reference of an engineer, scientist, or whomever, to make such attributions. This was the first consequence of those systems, like watches, cannons, and now AI, which do not define themselves.

The second matter concerns a distinction between "autonomy" and "heteronomy". What the notion of "closure of autonomy" reveals is that really these exogenously defined systems are *heteronomous*.

> "*Autonomy* and *heteronomy* literally mean, respectively, self-governed and other-governed. A heteronomous system is one whose organization is defined by input-output information flow and external mechanisms of control. Traditional computational systems, cognitivist or connectionist[71], are heteronomous. For instance, a typical connectionist network has an input layer and an output layer; the inputs are initially assigned by the observer outside the system; and output performance is evaluated in relation to an externally imposed task. An autonomous system, however, is defined by its endogenous, self-organizing and self-controlling dynamics, does not have inputs and outputs in the

[71] Whilst Thompson (2007) and others differentiate cognitivism and connectionism, as I discussed in the cognitivism chapter, they share the same fundamental concepts of computation, information, and functionalism and problem-solving conception of cognition. In this regard I have treated them as different expressions of the same paradigm of thought about mind and AI.

> usual sense, and determines the cognitive domain in which it operates…" (Thompson 2007: 43)

The inputs to which the Roomba robot is sensitive are "assigned by the observer outside the system" when those observers choose how to build it. The robot's task is "externally imposed", as are the metrics used to evaluate its performance at that task. Repeating it again, I have made the operative definition of meaning for an agent in this thesis "the *endogenously generated* significance an agent encounters in something". Naturally, the world the Roomba robot encounters takes on particular significance to it. However, the important thing is that the Roomba robot is not itself responsible as an agent for defining the way in which it interprets what it encounters. The significance for the robot of what it encounters is exogenously generated in the sense that the designers and builders of the robot specify how to respond to given features and stimuli. As Thompson says in the quote above, for example, "the inputs are initially assigned by an observer outside the system; and the output performance is evaluated in relation to an externally imposed task" (2007:43). This is not the case with autopoietic individuals. The basis and norms according to which autopoietic individuals respond to the world are endogenously generated. By contrast, both the robot's very constitution, and the coordinate system, figuratively speaking, which define what counts as meaningful functioning or malfunctioning are "other-governed".

To complement the difference between the self-governance of autonomy and the other-governance of heteronomy, consider now the difference between autopoiesis and allopoiesis:

> "…systems that do not produce themselves, but whose product is different from themselves, are said to be *allopoietic*. For example, a ribosome (a small spherical body within a living cell composed of RNA and protein, and the site of protein synthesis) is a crucial participant in the autopoiesis of a cell, but is produced by processes other than those that constitute its own operation (Varela, Maturana, and Uribe 1974: 188-19). Maturana and Varela also distinguish autopoietic systems from *heteropoietic* ones, which are allopoietic that arise in the realm of human design, such as cars and digital computers." (Thompson 2007: 98)

Heteropoiesis is a subset of allopoiesis. Both involve "other-production"(Maturana and Varela 1970: 80). The watch, the cannon, the robot, and Turing machines too, do not produce themselves, but something else. Their "work" and output produce something outside of, and other, to themselves. At this point the question of whether there is a legitimate self of which to speak, or whether such determinations have come round simply as a means of enabling sensible conversation about objects whose wholeness "lies outside the watch", is a growing elephant in the room. With that caveat, I will continue to refer to these heteronomous, allopoietic things as if there was a legitimate "there" there. In the case of the robot and a Turing machine, the output is a little clearer. A cleaned room and something printed or erased or moved or so on, on a "tape". For its ambiguity, the significance of the output being other than itself in the case of the watch and cannon reveals something important. The output of the watch is measured, based on its function, in terms of its capacity to "tell the time". This output is of significance only to an external agent. The same goes for the cannon and cannonball whose output is some destruction which is relevant and meaningful to an external observer.

For the purposes of this thesis, "internal norms of self-regulation" therefore mark a line between those systems and individuals which autonomously and autopoietically produce themselves, and for which the effort of their "work" is their very selves[72], and those heteronomous, heteropoietic, systems which are regulated and other-governed by external norms.

In this context of *internal* norms of self-regulation, I want to make a quick addition and commentary that situates autopoiesis within a broader context and brings in the kind of self-awareness unique to the Enactive paradigm and the legacy of 2nd Order Cybernetics whence it came (Varela et al. 2016, von Foerster 2003).

Autopoiesis concerns the dynamics of an individual and, as far as humans go, though, autopoiesis is not our only concern, even if it is the basis of our condition of being as physical organisms. It is decidedly individualistic. As long as we have been collective creatures, doing things for "others" has also been meaningful, and has even contributed to our capacity to autopoietically maintain ourselves. Donna Haraway, the author whose quote "It matters what thoughts think thoughts…what stories tell stories" inspires the spirit of this thesis, speaks of "sympoiesis" (Haraway 2016). In the third chapter of *Staying with the Trouble: Making Kin in the Chthulucene*, and acknowledging a heritage in (Margulis 1998), she presents sympoiesis as an ecologically sensitive update to autopoiesis:

> "*Sympoiesis* is a simple word; it means "making-with." Nothing makes itself; nothing is really autopoietic or self-organizing. In the words of the Inupiat computer "world game," earthlings are *never alone*. That is the radical implication of sympoiesis. *Sympoiesis* is a word proper to complex, dynamic, responsive, situated, historical systems. It is a word for worlding-with, in company. Sympoiesis enfolds autopoiesis and generatively unfurls and extends it… Lynn Margulis knew a great deal about "the intimacy of strangers," a phrase she proposed to describe the most fundamental practices of critters becoming-with each other at every node of intra-action in earth history." (Haraway 2016: 59-60) (emphasis in original)

The language and style are a bit different, influenced by anthropologies, ethnographies, and a posthumanist orientation that have their basis outside philosophy of AI. The significance of sympoiesis vis-à-vis autopoiesis is that autopoiesis occurs in a context. Much as autopoiesis enacts constraint-closure, it obviously occurs in the context of larger constraints, and with other autopoietic "critters" and "strangers", which means that autopoietic becoming is always a becoming-*with*. See (Barad 2007[73]), (de la Bellacasa 2017), and (Seibt 2020) for more work in this direction. A fuller account of post-cognitivism would include a discussion of sympoiesis, (and posthumanism), but autopoiesis is sufficient for the purposes of this thesis.

There is a line of thought here that is worth raising. It begins with the way in which, as long as humans have been sympoietically part of a collective, "work" has been done for some "other". This is of significance because, with a phenomenon like division of labour the work which a human enacts isn't

[72] The work of biological systems is not necessarily *entirely* devoted to self-production. This thesis is not a product of the biological work of my self-production, but then again, the work of humans is notable in this way for not being restricted to biological work. That is, we produce things like art, literature, psychology and endless other things which are not captured or explained solely in terms of the dynamics of biological work. This is one of the limits of autopoietic accounts of enactivism, addressed in works by (Di Paolo et al. 2018, Di Paolo et al. 2010.)
[73] Haraway's use of "intra-action" is a reference to Barad's 2007 work in particular.

entirely according to endogenous norms and, in this regard, something of the closure is lost. The "circle" of the "work cycle" is divided up into smaller arcs as work is no longer exclusively done for self-production but, and however admirably, for something "larger than itself". Now, the notion of "work" here is too ambivalent and I want to avoid possible misinterpretations: the biological work of autopoiesis, which as humans we enact as living organisms, is not identical to the work we do in an economic, division-of-labour context to put food on the table. Both are 'work', but one concerns biological enaction and the other *labour* in an economic market. Nonetheless, both are indeed a matter of work in a broad and appropriate sense because both involve the expression of energy through particular degrees of freedom, constrained through particular degrees of freedom. The difference is that the constraints in autopoietic enaction are endogenous whilst the constraints, often, in the case of economic labour are exogenous. In this way, the individual is no longer defining their own constraints necessary for their own enaction and so they are not entirely their own circle, not entirely an endogenously generated whole. Instead, the individual becomes in part an arc of a larger circle, specialised as a part of the larger whole. The meaning of the (economic) work of the individual has become decoupled from the individual's survival. The meaning of their work becomes *instrumental* to them in this way. Rather than Kant's "for, and by means of, themselves", there is "for an other, by means of themselves".

There is a clear similarity between this description of the long-time manner in which human collectives have socioeconomically organised, and the cognitivist conception of mind. Computation, information, and function concern the execution of goal-independent "work", work that is decoupled – orthogonal – to its meaning for a labourer. The work does not continuously contribute to the generation or maintenance of that which is produces the work. The significance of the work is for an external source, and so measured and determined by an external source. The work of a machine does not contribute to its own possibility in the way of autopoiesis. There is no constraint-closure in the work of a machine. Constraints are part of what is already-given in the cognitivist model of computation, information, and function.

I want to be clear. I am not suggesting that because of sympoiesis or division of labour that humans are not in fact autopoietic. I continue to maintain that, as *living organisms* we are indeed autopoietic and also sympoietic. However, where our *lived experience* in our social world is concerned, the way in which we as individuals participate in the meaning structures of society seems not so much autopoietic as computational. Again, there is a striking connection between the compartmentalised functionalism of the cognitivist conception of the organisation of cognition, and the division-of-labour socioeconomic organisation of human society. As living systems, our dominant model of mind are the principles of non-living, machines systems. Our mode of collective organisation appears to be based on these same principles, principles for building machines, not supporting the flourishing and fulfilment of living systems.

This brings forth an interesting proposition, namely, that we are autopoietic creatures living in computational social structures of meaning. It is hard not to imagine that the tension here may have something to say about experience of alienation characteristic for many modern humans in which many do not experience for themselves fulfilment in the work they do. This is to say, enactive autopoietic models may have something normative or action-guiding about them when it comes to the simple philosophical question of how to live, and how to live together. And this consideration is hardly independent from the purposes of this thesis either because, as we will see in the final chapters, the conclusions of this thesis are only about the different directions in which each paradigm leads us and do not offer guidance on which direction to take. The role of autopoiesis in the enactive model may be

able to offer something here. The thesis does not depend on this but, to the end of appreciating the direction in which this particular paradigm can lead us, it is worth consideration. In any case, the distinction between the autopoiesis and allopoiesis, therefore, and the distinction between endogenous and exogenous norms of self-regulation, speak to broader concerns beyond the particulars of this thesis and merit development.

Returning to the thesis, for now what matters is that this distinction between autopoiesis and allopoiesis is key to model of enactive autopoiesis. This is because autopoietic systems which define themselves also therefore define meaning on their own terms. Heteropoietic and allopoietic neither define and produce themselves, nor define meaning on their own terms. Rather than "internal norms of self-regulation", they are organised according to "external norms of exo-regulation". This is the case with the computational systems discussed in this thesis.

For such a system, its existence is not its own in the sense that the meaning of what it does is not something it defines or determines for itself. Whilst it produces its own output, the meaning of the output is not its own. The output is measured from a 3rd person perspective, by 3rd person metrics, much as Shannon measured the performance of electrical communication systems and we measure the performance of AI on arbitrary tasks. An AI system like this does not determine what is valuable or meaningful "for and by means of itself". Further, what is valuable is measured in a frame of reference not of its own determination, and according to a measuring stick that, too, is relevant and meaningful not to it, but to an external observer.

So far then, in discussing autopoiesis, it has been described in terms of self-producing wholes, and as complex systems which enact a closure of constraints such that their thermodynamic work goes back into the production of themselves. These dynamics of self-producing and self-defining closure set up a distinction be internal and external dynamics, endogeneity and exogeneity. In this section I have expanded on the significance of this distinction, comparing autonomy with heteronomy, autopoiesis with allopoiesis and heteropoiesis. What comes out of this comparison is a suggestion that, as autonomous systems, autopoietic individuals are self-governing, and define themselves "for and by means of" themselves. More broadly then, autonomous, autopoietic individuals *endogenously* determine what is meaningful for themselves, on their own terms.

It is to this suggestion which I will now turn in this coming, last section on autopoiesis. I will try to draw a connection between the condition of autopoietic systems as autonomous and endogenously self-producing and regulating, and how that condition is one in which a capacity of discernment of meaning and relevance happens. It is the final step from talking about autopoiesis to "sensemaking", the next concept with which I am endeavouring to characterise a paradigmatically post-cognitivist conception of mind. It is a shift from talking about individuals which *define themselves* "for and by means of" themselves, to individuals *which define meaning and relevance* "for and by means of" themselves.

5.2.2.5. Meaningful Perspective: Being a Frame

The point in this final section on autopoiesis is that autopoietic individuals *are* a frame of reference. It is admittedly as ambiguous as it is simple. Much as the frame of a picture effects a boundary which focuses the attention of an observer on a particular part or feature of the landscape in which it is embedded, so does an autopoietic individual enact a frame through which it looks a particular and relevant parts or features of the world in which it is embedded. It is an aperture. As an autopoietic

individual defines itself, it defines a condition through which – *as* which – it acts in the world. The condition of autopoietic being makes it sensitive to particular features of the world, those relevant to continued self-production. The condition of autopoietic being is therefore no less than a frame of reference in which, as which, and through which meaning *relative and particular to it* is defined.

More can be said about "being a frame". It can be understood in terms of the condition of "meaningful perspective" which autopoietic individuals enact. I will further suggest that autopoiesis is a *condition of observation* which necessarily generates particular observations, that is, observations particular to the condition of observation. To make it mildly more concrete, let us return to the example of the autopoietic cell producing and regulating itself, discussing this time the significance of the process of metabolism. Complementing the thermodynamic far-from-equilibrium discussion above, it is important for both understanding the condition of precarity of autopoietic individuals and, from that condition of observation, the way in which they must discern and define what they perceive in the world. Basically, being a metabolic creature means being a thermodynamically unstable creature, which makes certain things important to an organism.

This point will set up the discussion of sensemaking to come, because in the model of cognition as "sensemaking" the condition of observation makes all the difference. As I have emphasised so far in this discussion of autopoiesis, definition occurs on terms endogenous to the autopoietic individual. This applies to definition of both the self (autopoiesis), and discernment of the world (sensemaking).

As far as "being a frame" goes, and understanding what a Frame is, here is a basic recap from the third chapter in which I presented the Frame problem as a fundamental theoretical problem of cognitivism. It has a more narrow conception in the domain of logic, concerning defining and keeping track of legitimate inferences (Shanahan 1997, 2004) and a broader philosophical conception concerning the question of how to specify in a non-arbitrary way what counts as relevant for an agent to consider. Much of the thinking here is intuitive – a frame of reference is a line which defines the boundaries of the picture. That boundary marks the edge of what is included from what is excluded. It defines a domain of awareness and consideration. We do not perceive what is excluded, but we know there is more to see than is shown. Of all the possible place to look, "look here". Much as the constraints in the earlier context of complexity define degrees of freedom through which energy is expressed as work, a frame of reference is a set of constraints which define degrees of freedom through which awareness is expressed. Further, inasmuch as the closure of autopoietic complex systems means they define their own constraints which produce themselves, so do they define their frame of reference, their particular condition of observation. This is to say, their *perspective*. A frame of reference is a condition of observation is a perspective. A frame de-fines - renders finite. A framed perspective in this context is then this difference between all-seeing view-from-nowhere omniscience and the "all-too-human" bounded perspective.

In the way that it includes some things and excludes others, a frame of reference defines what is significant and relevant in that perspective. A condition of observation enables observation of certain things and not others. Different conditions of observation are required to observe different things. Telescopes and microscopes "extend" our optical conditions of observation such that we can see the very small and the very far away. In both cases something is included, and other things are excluded. In a more abstract way, theories are also conditions of observations. Feminist philosophies of science (e.g.: Barad 2007) and critiques of philosophy have shown how theories are generative of perspectives which can be exclusive in ways that are profoundly unjust (e.g.: Fricker 2007). This is to acknowledge and anticipate a coming discussion in the next chapter on the *moral* significance of frame. It is a position of power to draw that line that includes and excludes, defining what is to matter.

To say, therefore, that a frame defines by inclusion and exclusion what is significant and relevant – what is meaningful – and to say that autopoiesis involves the self-production of a frame and perspective, means that *autopoietic systems define their own condition of meaningful perspective*. They define their own conditions of observation and therefore, in this way and to this extent, define a condition in which what is meaningful for them, is defined by them, "for and by means of themselves". To wield the weight of "closure" again, we might say there is "closure of framing". Autopoietic individuals produce themselves, producing a frame of reference which defines certain things as relevant, relevant to the continued production of that frame of reference[74]. Compare this endogenous production of frame with the discussion of the previous section. There I spoke about the significance of the endogeneity of "internal norms of regulation", and about how, in contrast, heteronomous and allopoietic systems define neither the tasks which they undertake, nor the frames of reference in which performance at those tasks is measured. Of course, those perspectives may exist because external observers can still observe an autopoietic organism, they just do not define what is meaningful to the organism. Their measurements are not the relevant measurements for an autopoietic system. They *are* for an allopoietic system[75].

A "closure of framing" means that autopoietic systems endogenously define the condition of their own perspective, and therefore also the de-fined domain of what counts as significant for them. In this respect, autopoiesis can be understood as self-production of *meaningful perspective*, a framed condition of observation with a defined boundary including what is relevant and excluding what is not, by and from the perspective of the autopoietic organism. This is not something which contemporary computer systems and AI systems, as exogenously generated systems, can do.

To make this mildly more concrete, this is often spoke of in terms of the dynamics of metabolism in biological and ecological systems (Stewart et al. 2010, Thompson 2007, Froese and Ziemke 2009). In its strict sense, metabolism refers to the chemical processes which sustain life, converting energy in food to energy for cellular processes as well as the conversion food into building blocks for these processes, and excretion of waste products. More generally, metabolism might be understood as the appropriation, digestion and transformation of something once external to support internal process. In this way metabolism becomes a metaphor for integrating anything, from physical things like food, to abstract things like ideas and psychological experiences. Insofar as it is a discernment across a semi-permeable threshold, metabolism is a close analogy to cognition. What is significant is the discernment involved in the metabolic process. In the transformational process of metabolism, what is supportive and conducive to life is metabolised, and what is waste is excreted. Choice amongst possible alternatives" is enacted. The transformation process of metabolism is the continuous transformational process of the autopoietic individual, continuously reaffirming – enacting – itself. In a passage I cited partially earlier, Thompson says:

[74] "Continued frame of reference" does not mean that an organism has or is an identical frame of reference through the course of their lives. Any change in the organism means a change in the frame as which it is perceiving the world. See just above this footnote, at the bottom of this page, the discussion of "continuous reaffirmation". Again, this will make more sense once the notion of sensemaking is involved. For the intuitions - what is relevant when we're hungry, and what the world looks like, is different to when we're stuffed. The same applies to larger developmental differences an organism undergoes - what is meaningful to a child relative to an adult.

[75] This distinction may have something important to contribute to the conundrums of Searle's (1980) "Chinese Room" thought experiment and Dennett's summarising quip about there being "competence without comprehension". What the man in the room is doing is not meaningful to him. He doesn't understand what he is processing. The significance lies with some external party. Perhaps someone who actually speaks the language and who can arbitrate his performance, otherwise it is mostly significant to external philosophers trying to make sense of it.

> "A lifeless thing does not metabolize; hence "its duration is mere remaining, not reaffirmation" (Jonas 1966: 81) …The organism must eat and excrete; otherwise it dies. Without incessant metabolic exchange with the world there can be no emancipation of dynamic selfhood from mere material persistence…
>
> Metabolism is the constant regeneration of an island of form amidst a sea of matter and energy. *Metabolism establishes a self with an internal identity* marked off from the outside world and whose being is its own doing. *Metabolism operates according to internal norms that determine whether otherwise neutral events are good or bad for the continuation of the organism*.
>
> An organism must subordinate every change it undergoes to the maintenance of its identity and regulate itself and its interactions according to the internal norms of its activity. Life is thus a self-affirming process that brings forth or enacts its own identity and makes sense of the world from the perspective of that identity. The organism's "concern", its "natural purpose," is to keep on going, to continue living, to affirm and reaffirm itself in the face of imminent not-being." (Thompson 2007: 152-153) (Emphases my own)

In establishing a "self with an internal identity", metabolism *endogenously* established a condition of observation, a frame of reference in which "otherwise neutral events are good or bad for the continuation of the organism" (ibid). Metabolism is a concrete expression of meaningful perspective, of "being a frame".

The endogeneity of this process matters. As I discussed in the previous section, things which a computer or Turing machine encounter will take on specific significance for them. Such systems have a perspective, but there is an instrumentality to the perspective. The perspective is exogenously generated by whoever built the system. Repeating the theme, autopoietic entities endogenously generate their own condition of meaningful perspective, but computers do not.

To summarise this section, then, particular conditions of observation generate particular observations. Particular frames of reference define particular domains of perception. Autopoiesis is the self-production of a condition of observation, a frame of reference, and hence observations unique to the frame and condition of the individual. This is significant in the frame of this thesis to the extent that it supports a characterisation of post-cognitivist conception of mind in terms of "problem-defining". Again, what I mean by "problem-defining" is not an activity in some strict sense in which we articulate some particular problem, but rather the more fundamental orientation an individual endogenously establishes, and, on which basis they discern meaning in their world of perceptions. To say of autopoietic systems then that they "are a frame" then is to say that they are a condition of observation, one of meaningful perspective. It means that, insofar as they produce and frame and define the degrees of freedom of their awareness, they are defining the context in which anything can be measured as a "problem" to solve.

In the next section on sensemaking, I will move from this higher order discussion about frames and the condition of observation to talk about the content that is perceived. Naturally, it too is agent-relative in this fundamental way. Between the section on sensemaking and the final section on dynamic co-emergence, I will talk about what might well be a slogan of enactive and post-cognitive conceptions of mind - that "being defines a domain of relevance". This slogan of sorts can be read as expressing the idea that we perceive as a function of the kinds of beings, i.e. conditions of observation, we are. That is, that we define the world and its contents in ways that are relevant to us. If this doesn't support a

characterisation of post-cognitivist conceptions of mind in terms of problem-defining, I don't know what could.

5.2.3. Sensemaking

> We do not see the world as it is, we see it as we are.
> - Talmudic saying
>
> If a lion could talk, we could not understand him.
> - Wittgenstein

Moving from autopoiesis to the notion of sensemaking, we move from talking about a condition of being to a mode of cognition particular to that condition of being. It is a mode of discernment of an autopoietic self[76]. Thompson defines it as "…the exercise of skilful know-how in situated and embodied action" (2007: 13) This is notably similar to Dreyfus' Heideggerian "coping" (Dreyfus 2007) but not all that illuminating for anyone not already versed in these languages. The point is that meaning, significance, and relevance for the agent are all inherent to sense-making the agent enacts because in order to maintain itself, an autopoietic organism must have a capacity to discern what is relevant in that regard; sense-making is "meaning-generation in relation to the perspective of the agent...." (Froese and Ziemke 2009: 481).

It is on this understanding of cognition that post-cognitivism is being characterised in this thesis in terms of problem-defining. As a reminder, recall that problem-defining is an individual's "endogenously generated orientation and discernment of meaning" in their situation (§5.1.). In the end, sensemaking is the detailed and technical account of enactive cognition, of which "problem-defining" is a basic rendering.

Sensemaking can therefore also be thought of as "endogenous orientation and discernment of meaning", but, as it concerns living systems (the life-mind continuity thesis), sensemaking has an important biological sense to it which highlights *adaptivity* (Di Paolo et al. 2017, Di Paolo 2018). To integrate this, we can think of sensemaking as endogenous *adaptive discernment*, discernment of those "differences that make a difference" in both self and environment, relevant to the integrity and survival of the autopoietic self. In the condition of a precarious, semi-permeable identity, it is adaptive to be able to discern what is relevant, internally, and externally, to one's survival.

Just like with problem-defining, sensemaking individuals can be "wrong" in their adaptive discernment of meaning. That is, there can be a "mismatch" between what an organism discerns for itself to be adaptive and what, from an external perspective, we would say is in fact adaptive for it. So, whilst I am suggesting that cognition as sensemaking can be understood as adaptive discernment, adaptivity is not always guaranteed.

Sensemaking as adaptive discernment has an important implication, namely, the discernment is *perspectival*, or relative to the perspective of the organism. This is the same "meaningful perspective"

[76] It therefore applies to any autopoietic system. Recall from the discussion in §5.2.2.1. on Self-Producing Wholes which, according to Thompson's three criteria (semi-permeable boundary; reaction network; and interdependence of both of these things), "a bacterium (a prokaryote) and an amoeba (a eukaryote) are autopoietic because they satisfy all three criteria" (Thompson 2007: 103). However, whilst autopoiesis is generally recognised as necessary for cognition, there is some debate as to whether it is sufficient. I will discuss this momentarily.

discussed earlier in the context of autopoiesis (§5.2.2.5.). We need not say that sensemaking involve a commitment to a full blown "mind-dependency" picture though. Sensemaking as adaptive discernment is the mode of cognition of autopoietic, living systems which means that sensemaking includes microfauna like bacteria and amoebas, creatures for which the suggestion that they have a mind might be controversial. It is sufficient to say that the endogenous discernment is from the perspective of the organism, given its particular capacities of perception and that, as such, just like with problem-defining, features discerned endogenously take on specific meaning for an organism that may be different to another organism which might not even discern the feature.

There is a further implication to note before continuing. If the adaptive discernment is relative to the perspective of the organism, we can ask about whether what the organism discerns is metaphysically real or, more generally, whether this enactive theory of cognition takes a position of non-realism or something similar. Realism questions for the enactive theory of mind and cognition do not appear to be settled. For one position on the matter, John Stewart, in a recent textbook on enactivism, asserts that enactivism is "radically constructivist" (Stewart 2010: 27).

One way of talking about it that may avoid realism concerns is to consider again that particular conditions of observation generate particular observations, and autopoiesis is the self-generation of meaningful perspective (§5.2.2.5), meaning, autopoietic entities will observe only what they can. This stands in contrast to the perspective-less, or "Frame-less" "view from nowhere" expressed in conceptions of computation which articulate a picture of cognition without perspective – Dupuy speaks of this in *The Mechanization of the Mind* as "subjectless cognition" (2000: 156).

Having now suggested that sensemaking can be understood with certain caveats as adaptive discernment, a more biological sense of "problem-defining", we can return with this understanding to a discussion of sensemaking as it is understood in the literature on the enactive theory of mind and cognition. I will begin with a return to the life-mind continuity thesis because, for early thinking on enactive cognition, "living is sensemaking" (Varela 1991, 1997, Weber and Varela 2002). This claim lends itself to the idea that, as living systems, "autopoiesis too is sensemaking". However, it should again be noted that whilst autopoiesis is necessary for enaction realisation of an organism, it is not universally accepted as sufficient for cognition (Thompson 2007: 122-127). Saying a bit more about this will help clarify what is meant by cognition in the context of autopoiesis and thus what sensemaking really amounts to. With an understanding of what makes an autopoietic system autopoietic in place, I can begin to consider what the cognitive process looks like. Here I will discuss the "sensorimotor loop", the way living systems make sense of themselves and the world. From there, I will note the role of adaptivity for sensemaking individuals. With all that in place, a definition of sensemaking will then make more sense. In particular, I want to frame sensemaking as an endogenous capacity for discernment of those meaningful differences, or in the terms of this thesis, defining problems and features of the world in ways that are relevant. Sensemaking then is the next key concept in establishing the "problem-defining" character the post-cognitivist paradigm of mind.

5.2.3.1. Living is Sensemaking

To bridge the connection between autopoiesis and sensemaking, we can return to the life-mind continuity thesis, for, "living is a process of sense-making, of bringing forth significance and value" for the organism (Thompson 2007: 158). In the earlier discussion, life-mind continuity established that the dynamics of living systems were sufficient to account for cognition generally. Now we are looking

more specifically at enactive cognition, that is, sensemaking. To unpack the identity claim here, it is helpful to clarify what the differences being bridged are. Thompson has a helpful distinction between the autopoietic dynamics of the living, and cognition: "Autopoiesis" pertains to the self-producing organization of a living system, whereas "cognition" pertains to the behaviour or conduct of a system in relation to its environment." (ibid: 124)

Note that this conception is consistent with a basic cognitivist conception of cognition, like Russell's who, speaking of intelligence, offers this initial description - he says it is to be found "…in a simple relationship between what we perceive, what we want, and what do. Roughly speaking, an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived." (Russell 2019: 14) There is no mention of environment as such, and nothing like the "closure of autonomy" and "closure of framing". So, there are differences, but otherwise the conception of cognition here in relation to autopoiesis is pretty basic.

Returning to the question of whether autopoiesis necessarily entails cognition, there exist some very basic systems that exhibit what could be considered "minimal" autopoiesis or "proto-autopoiesis", but which do not interact with their environment, and so do not count as cognitive in this sense. Bourgine and Stewart (2004) speak about a digital "three-dimensional tessellation automaton" they consider to be minimally autopoietic, but not cognitive. There are also examples of small autocatalytic biological systems which autopoietically produce their own boundary, but do not actively engage their environment and so also can be said to be autopoietic but not cognitive (Bitbol and Luisi 2005).

Thompson summarises in stating that, if autopoiesis is understood as only meaning endogenous self-production and does not include interaction with environment, then autopoiesis is not sufficient for cognition, but if autopoiesis does include interaction with the environment, then it does entail cognition (Thompson 2007: 126-127). Following Thompson, it is enough that a system is living, for then it is both autopoietic and cognitive in the sense discussed above.

5.2.3.2. Sensemaking: The Sensorimotor Loop

With that caveat, the next step for building up to the notion of sensemaking is the role of the "sensorimotor loop" in living, biological and autopoietic systems (Newen et al. 2018: §4-5). These systems, cellular or animal, are embodied. As these autopoietic bodies endogenously produce and enact themselves, they make a difference in the world. This is as simple as the notion of displacement, the way Archimedes' body displaced water in the bathtub. The body makes a difference to water. Simultaneously, the water exerts pressure on his body. There is "mutual-difference" making. This notion is central to understanding how the sensorimotor-loop of an autopoietic body is involved in sensemaking process of cognition.

As an autopoietic organism produces and maintains itself, it makes a difference to the world sheerly by "taking-place" - in French, to say something happened is to say it "took place" – dis-placing, known, defined, and individuated by the difference it makes. The organism makes a difference to the features of the environment in which it is embedded, much as a body makes a difference to the water in a bathtub. The environment puts pressures on the body, the body produces itself, holding itself both ordered enough to maintain itself, and open enough to allow a metabolism of energy, matter, and nutrients so that it can continue to produce itself.

Either way, a sensorimotor loop happens then when the body, or actuators (read: "difference-makers"), effect a difference in the world, and sense the world in the way in which, and to the extent that, the world "pushes back" on the actuators. In response to the feedback, the individual coordinates itself, which in turn effects a difference to which there is again a mutual difference-making feedback, and so on. There are several important implications to this "sensorimotor coupling" (Froese and Ziemke 2009: 471) between agent and world.

The first is the relationship between perception and action. In his book-length treatment of this matter, Alva Noë talks about "action as perception" (2004). It is in the motor activity of a body that the senses are stimulated, and according to those senses that the motor activity of the body is coordinated.

The second implication is the way in which the enactive account of cognition operates, however implicitly, via the metaphor of "touch" (Noë 2004, Ratcliffe 2018). This is in contrast to the operative metaphor of vision, again, however implicit, of cognitivism (Marr 2010). A prominent difference between the cognitive, and also epistemological, processes understood in terms of these two senses is that mutual-difference making is not a feature of a vision-based computational model of cognition. Recall in Russell's notion of cognition there was no explicit mention of the environment, let alone the way it makes a difference to an agent.

Thinking with the metaphor of vision, intuitions suggest it is possible to see without being seen, know without being known, and make a difference without suffering it, that is, without being changed. There is a separation of knower and known, observer and observed. The epistemic process is asymmetric. Whilst it is possible – easy – to imagine that objects exist when I do not look at them, it is not possible to touch without being touched. It is not possible to make a difference without cost. There is no free cognitive or epistemological "lunch" in this respect. It is a symmetrical epistemic dynamic. Much as trying to grasp water displaces the very thing we're trying to grasp, it is not possible to measure without making a difference to the very thing being measured. To measure it is to make a difference to it[77]. Thinking about cognition and epistemology with the metaphor of touch, there is no clean lens mediating the interaction between self and world, isolating the "controlled" system under observation. There is no innocence, only muddy, fleshy participation.

Despite this novelty, the notion of a sensorimotor loop has long been adopted, even in cognitivist spaces, and notably in the robotics space (Brooks 1990, 1991, 2002). Rodney Brooks has a famous line that "the world is its own best model" (ibid 1990). It is a response to the challenge of trying to define the environment for a cognitivist agent in the form of a representational model, because such a model assumes the relevantly pre-defined features of its environment. Echoing Russell earlier, Brooks asks: "but what is the correct symbolic description of the world around the intelligence system? Surely that description must be task dependent." (ibid: 4) Rather than computing a representational model of the world, Brooks' advance was to provide robotic systems (which I include under the banner of "AI") with sensors which generate real-time information about the environment.

[77] This relates to the "measurement problem" of quantum physics, which would take things too far afield to integrate here. Philosophically, the measurement problem is the question of why measuring something seems to affect the measured value. This is not a problem in classical physics with objects of macro size – as long as we use the same measuring stick, measuring the length of this laptop will give the same measurement each time. With macro-objects, the "length" seems obviously independent of the measurement. In the case of quantum scale phenomena, "measurement" does not have this kind of insulation from the measured thing. "Measurement" makes the same difference as an interaction. Famous physicists Werner Heisenberg and Niels Bohr have articulated epistemic (Heisenberg) and ontological (Bohr) accounts of what is going on here (Barad 2007). These discussions connect with the Realism questions of sensemaking accounts of cognition – particular conditions of observation generate particular observations.

> The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough. (Brooks 1990: 5)

In contrast to a conventional Turing machine type system whose (exogenously defined) inputs and outputs have meaning to the extent that they represent things in the world for the external observer, a system embodied with sensorimotor faculties begins to process cues "grounded" in the physical work. Froese and Ziemke (2009) comment on the significance of this development in the context of development of AI:

> The focus on the organization of sensorimotor situatedness has several important advantages. The crucial point is that it enables an artificial agent to dynamically structure its own sensory inputs through its ongoing interaction with the environment. Such a situated agent does not encounter the frame problem, more generally conceived, because of its tight sensorimotor coupling with the world. It never has to refer to an internal representation of the world that would always quickly get out of date as its current situation and the world around it continually changed. Furthermore, it has been claimed that for such an agent "the symbol grounding problem is really not an issue – anything the agent does will be grounded in its sensory-motor coordination". In other words, from the perspective of embodied AI it seems that the problem of grounding meaning has been practically resolved by generating artificial agents that are embedded in their environment through their sensorimotor capabilities. (Froese and Ziemke 2009: 471)

5.2.3.3. Sensemaking: Constitutive Autonomy and Adaptivity

As a design principle for artificial systems, a sensorimotor loop clearly brings a lot, then, but the same developments in robotics reveal limits that suggest sensorimotor coupling is not sufficient for "meaningful perspective", that is, the endogenous self-production of perspective (ibid: §2.3). Froese and Ziemke identify two necessary conditions for sensemaking, "constitutive autonomy" and adaptivity (ibid: §3.2) (See also (Thompson 2007: 148).

"Constitutive autonomy" is no more than the "closure of autonomy" of autopoietic systems, which has already been discussed. Again, it is the way in which autopoietic systems are responsible for their own constitution in a way that a tool, a car, a cannonball, and a watch are not, despite being autonomous once they get going. Adaptivity pertains to the meta-stability of an autopoietic organism as it endeavours to maintain itself in the course of the mutual difference-making of its sensorimotor activity in the world. As it moves in the world, the world mutually moves it, and so it must adapt if it is to remain an individual.

I will note why sensorimotor loops are insufficient for "meaningful perspective" and then explain how constitutive autonomy and adaptivity are necessary for sensemaking, tying together "meaningful perspective" and sensemaking.

The key point about sensorimotor loops is that the sensory information must be meaningful to the agent, not an external observer. Whilst sensorimotor coupling with environment does involve cognition insofar as it pertains to a feedback loop between agent and environment, for it to be sensemaking, the meaning of the stimuli must be relative to the agent and endogenously generated. The frame of reference must be endogenous to the agent, not externally defined by an external observer. For sensemaking, the meaning must be *endogenously* meaningful for the individual. Again, this is a key difference between cognition as computation and cognition as sensemaking. A machine whose output is meaningful only to an external observer can be said to be computing, but it is not sensemaking unless the meaning is endogenously generated. Consider the following example of Brooks' Roomba robot.

A Roomba robot has a sensorimotor coupling with its environment, navigating with more or less competence. What counts as meaningful behaviour of the robot is not endogenously determined by the robot or for the robot, but by and for an external observer. There is "competence without comprehension", to use Dennett's phrase, much as there is in Searle's "Chinese Room". In both cases, what is meaningful is, again, not defined endogenously, but exogenously. For example, according to this distinction we can say that whilst the agent in the Chinese room is computing, (and in this sense the Chinese symbols encountered obviously do take on a "particular significance" for the agent), the agent is not, however, *making sense* of what they are processing. The basis for their discernment is exogenous. It is specified by a book of rules for how to respond to given input[78] .

Now, it might be objected that this requirement of endogeneity unique to autopoietic sensemaking implies that only living things can have "meaningful perspective". For the enactivist this is not a problem. It is the starting position of the paradigm. It is what the life-mind continuity thesis ultimately means. "Living is sensemaking" (Maturana and Varela 1987).

This "living" is the "constitutive autonomy", the autopoiesis, the constraint-closure, "closure of autonomy" and "closure of framing". Because autopoietic systems endogenously produce and constitute themselves autonomously, they are endogenously responsible for maintaining themselves through the sensorimotor perturbations effected by the mutual-difference making of their sensorimotor coupling. This means that the differences living and autopoietic systems encounter make a difference to the integrity of themselves, that is, to their very being. Autopoiesis and self-organisation at thermodynamic far-from-equilibrium mean that the stimuli are interpreted in terms of how they contribute, positively or negatively, to the maintenance of that autopoietic identity. Thompson emphasises this point with the example of bacteria:

> "…the now familiar example of motile bacteria swimming uphill in a food gradient of sugar. The cells tumble about until they hit upon an orientation that increases their exposure to sugar, at which point they swim forward, up-gradient, toward the zone of greatest sugar concentration. The behaviour occurs because the bacteria are able to sense chemically the concentration of sugar in their local environment through molecular receptors in their membranes. They are able to move forward by rotating their flagella in coordination like a propeller. These bacteria are, of course, autopoietic. They also embody a dynamic sensorimotor loop: the way they move…depends on what they sense, and what they sense depends on how they move. *This sensorimotor loop both expresses and is*

[78] Note that this rule book is one of the already-given conditions of the "cognitive situation" in the thought experiment, which suggests that it makes cognitivist assumptions.

> *subordinated to the system's autonomy, to the maintenance of its autopoiesis.*" (Thompson 2007: 157) (emphasis mine).

This "subordination" of the sensorimotor loop to the maintenance of autopoiesis is the difference between, on the one hand, the way in which the successful function of a Roomba robot is meaningful to an external observer, but not to the robot itself, (because it is exogenous), and, on the other hand, the way in which the sensemaking of an autopoietic organism is *endogenously* meaningful to it. The difference, again, is that the robot is heteronomously, and thus exogenously, produced. It is not responsible for its own production. The autopoietic constitution of the organism is autonomous and endogenous. It is responsible for its production and maintenance and therefore endogenously makes sense of the world in a way that enables that. Constitutive autonomy is therefore necessary so that the sensorimotor loop is relative to the autopoietic individual, so that the meaning of the stimuli is endogenously meaningful *for it.* Froese and Ziemke press this point in their description of the phenomenon of sensemaking:

> Furthermore, what an autonomous system does, due to its precarious mode of identity, is to treat the perturbations it encounters from a perspective of significance which is not intrinsic to the encounters themselves. In other words, the meaning of an encounter is not determined by that encounter. Instead it is evaluated in relation to the ongoing maintenance of the self-constituted identity, and thereby acquires a meaning which is relative to the current situation of the agent and its needs. This process of meaning generation in relation to the perspective of the agent is what is meant by the notion of *sense-making."* (Froese and Ziemke 2009: 481)

The link from constitutive autonomy to adaptivity comes out in Kant's expression that I have used several times so far in this chapter – the "for and by means of itself" of autopoiesis. Autopoietic sensemaking is "for and by means of itself", *endogenously for* the autopoietic self, and *endogenously by means of* the autopoietic, bodily self. As Dreyfus notes from Merleau-Ponty, whose phenomenology is a major source of philosophical inspiration for the enactive theory, "through [my] body I am at grips with the world" (Dreyfus 2007, Merleau-Ponty 2002: 353). It is through, with, and "by means of" its autopoietic body that such an individual makes sense.

This means several important things. Perhaps the deepest is that there is no separation between knower and known, but rather an intimate entanglement of bodies mutually affecting each other. I cannot make sense of water in an embodied way without getting wet. In this way, sensemaking is "involved," and "participation" is not distanced or safe. There is no insulation, no system closed-off from what it is epistemologically meeting, but rather only those "semi-permeable" boundaries of thermodynamically open systems, regulating passage across thresholds. This processing across the threshold of the autopoietic individual is the cognitive process. In this way, cognition is the integration of difference to the very being of an autopoietic being, as it metabolises energy and matter. Differences do not just make

a difference to a model of an autopoietic organism[79], but to the enacted embodiment of the autopoietic individual[80].

In any case, endogenously making sense of the world "by means of" the body means being open enough and vulnerable enough to let it in, to get "wet". As the quip goes, "a model of a hurricane is not wet". They require very different adaptive responses. The adaptivity of sensemaking identifies that there is a risk and a stake inherent to sensemaking. Adaptivity is not always guaranteed. Sensemaking can go "wrong" in the sense that an organism can discern things in a way that are maladaptive.

Sensemaking is therefore endogenous meaning-generation relative to the perspective *at stake*. Sensorimotor loops without a stake, subordinate to which an organism must adapt, may be sufficient for the perspective of a robot, but it is not "meaningful perspective" without the stake and adaptivity.

There is more to say about stake. Whilst the external observer of a heteronomous system like a Roomba robot obviously has something to gain and lose from the functioning or malfunctioning of the robot, neither the observer's life nor the robot's are at stake in the way they are in autopoietic sensemaking. Of course, the context of this thesis – the ostensible existential risk of artificial superintelligence – is precisely a concern about the external observer, a concern that we *do* have skin in the game where the capacities of discernment of AI are concerned. This may be the case, but it doesn't mean that AI is sensemaking. It is still just executing an exogenous function. What it *might* mean is that we as external observers are sensemaking via the "extension" of a highly sophisticated tool, and that there is a possibility that the tool becomes a "loose cannon", putting humans at risk. Just because there is a high-stake for the external observer when it comes to artificial superintelligence, does not mean that the artificial superintelligence is sensemaking. Its frame of reference is still exogenously defined by an external observer as is the basis of any meaning it discerns. What "skin" AI has in the game is not its own.

The "skin in the game" of autopoietic organisms is very much their own, and nothing less than their life depends on the appropriate capacities of sensemaking. In the enactive model of cognition, staying alive depends on a capacity to discern and define what will contribute positively and what will contribute negatively. As Thompson puts it: "Something acquires meaning for an organism to the extent that it relates positively or negatively to the norm of the maintenance of the organism's integrity." (Thompson 2007: 70)

What is adaptive, what *matters,* what is *relevant*, is necessarily endogenous, relative to the organism. Sensemaking is navigation of a world unique to the perceptual capacities and modalities of an autopoietic organism. It doesn't mean we always perceive things that are adaptive, we obviously also perceive things in a way that is maladaptive, which is something this account has to contend with. But in any case, my dog has different sensorimotor capacities to me and thus is making sense of very different features. The features which are relevant and adaptive for her are going to be very different to those that are for me. In this way, she is making sense of a remarkably different world to me. It is to

[79] Enactive accounts and the post-cognitive paradigm generally do admit of abstraction though, see (Villalobos and Dewhurst 2017, Stewart 2010: 4).

[80] This brings to mind Frank Jackson's thought experiment about Mary, a colour scientist (Jackson 1982, 1986). It seems like the difference between the abstract propositions Mary learns about colour and the embodied, sensorimotor experience of colour is a difference that might be meaningful to Mary, much as getting wet and learning of it are different, stimulating different senses and, importantly, requiring different adaptive responses and sensemaking.

this difference that enactive "dynamic co-emergence" speaks, the last of the concepts of enactivism I will address.

Before turning to dynamic co-emergence, I want to summarise this discussion of sensemaking with an offering of a few different definitions. From there, sensemaking can be compared with the cognitivist notion of cognition as computation.

5.2.3.4. Summary of Sensemaking

So far, in this discussion of sensemaking, I have described several characteristics of sensemaking. I discussed the significance of life-mind continuity in the claim that "living is sensemaking", I discussed the significance of an endogenous sensorimotor loop, and just now finished a discussion about the role of "constitutive autonomy" and adaptivity. I want now to finish by presenting several definitions which collectively present a sense of sensemaking which connects with the problem-defining character of post-cognitivism. I follow with my own sense of sensemaking afterwards.

> "Living is sensemaking" (Varela 1991, 1997, Weber and Varela 2007, Thompson 2007)
>
> "Sense-making is the enaction of a meaningful world *for* the autonomous agent. (Froese and Ziemke 2009: 481)
>
> Sense-making as adaptive coupling between living body and world-environment. (Cappuccio and Froese 2014: 5)
>
> "Exchanges with the world are thus inherently significant for the agent, and this is the definitional property of a cognitive system: the creation and appreciation of meaning or *sense-making*, in short...It will be important to notice already…that sense-making is an inherently active idea. Organisms do not passively receive information from their environments, which they then translate into internal representations. Natural cognitive systems are simply not in the business of accessing their world in order to build accurate pictures of it. They participate in the generation of meaning through their bodies and action often engaging in transformational and not merely informational interactions; *they enact a world.*" (Di Paolo et al. 2010: 39) (emphasis in original)
>
> "For enactivism, value is simply an aspect of all sense-making, as sense-making is, at its root, the evaluation of the consequences of interaction for the conservation of identity." (ibid: 45)

Because the point of this chapter, and thesis, is not to defend a particular account of cognitivism or post-cognitivism, or sensemaking, but get a sense for the way of thinking in each, all of these accounts of sensemaking can be included and accepted. The chapter is aiming for a sense of the whole.

In chapter 3, I offered my own summary definition of computation as "rule-bound information-processing". The purpose was to highlight the significant aspects of computation which hold across the conceptions of cognitivism. The same can be done for sensemaking. Here are a few similar attempts:

> Sensemaking is *endogenous,* open-ended meaning-generation relative to the perspective *at stake.*

> Sensemaking is open-ended navigation of a world unique to the perceptual capacities and modalities of a self-producing and maintaining individual.
>
> Sensemaking is open-ended, endogenous adaptive discernment.
>
> Sensemaking is an endogenous capacity to discern what is going on, a capacity to define "the problem" in a way that is relevant, from the perspective of the organism.

The last definition is the least "technical". "An endogenous capacity to define *the* problem" may sound like the problem already exists, independently of the organism's perception or conception, but this is not the intended meaning. "The" problem is the problem endogenously discerned "for and by means of" the organism.

The definition of sensemaking as endogenous adaptive discernment is the simplest of the suggestions. It might be thought that endogenous adaptive discernment reduces sensemaking to evolutionary adaptiveness. This would be misleading because an organism can be "wrong" in their discernment insofar as they do things which can be described as maladaptive. Further, it is vital to an understanding of sensemaking that the discernment is endogenous and from the perspective of the individual in question. Questions of adaptivity can invite external or objective observations about what is in fact adaptive for the organism. Ultimately sensemaking is about endogenous meaning too. Sensemaking recognises that features of an organism's world endogenously take on specific meaning for the organism. External or 3rd person observations can make inferences about what is meaningful based on observations about what is adaptive, but the 1st person perspective must have a certain priority and authority in any modelling because that is the endogenous perspective.

Based on this line of thinking, the following summary definition of sensemaking can be offered:

> Cognition is sensemaking iff it is open-ended, endogenous adaptive discernment of meaning by an autopoietic organism, "for, and by means of, the self".

I have deliberately described sensemaking in terms of discernment so as to connect with the "problem-defining" character I am suggesting of enactivism and post-cognitivism. It is my hope that what has been said in this chapter so far offers a sense that this is not an arbitrary, or trivial, way of framing and making sense of sensemaking.

Before finishing, I want to bring in the two quotes with which I began this section on sensemaking. They each offer for our intuitions a simple and potent sense for sensemaking. The first is a saying the source of which I cannot be sure, but seems to be attributed to the Talmud, a central text of Rabbinic Judaism: we do not see the world as it is, we see it as we are. Then there is Wittgenstein's (in)famous quip that if a lion could talk, we could not understand him. Between Thomas Nagel's bat, Wittgenstein's lion, my dog, and a Roomba robot, "we do not see the world as it is, we see it as we are". What is relevant to us is a function of what and who we are, and what we are capable of seeing, given our unique conditions of being (Cannon 2022). As Stewart notes just above, "what the knowable world is, for each of us, is not independent of who we are" (Stewart 2010: 27) - sensemaking as "…navigation of a world unique to the perceptual capacities and modalities of a self-producing and maintaining individual." In this light, we could not understand Wittgenstein's lion because, perceiving the world as the lion is, the meaningful features of its world would be features we do not perceive, much as features which define

our worlds – "the market", "the internet", "AI" – are not features the lion would be capable of perceiving. There is something mildly paradoxical, puzzling at least, implied in this situation that we can all be in the same world and see different things, as if we both are and are not in the same world. I will pick this up again in the next section on "enactive dynamic co-emergence of self and world".

As a last roundup, to further support this framing, and to summarise this section on sensemaking, I will now compare it with the computational account of cognition and intelligence presented in the previous chapter. This should hopefully help make sense of the characterisations of cognitivism in terms of "problem solving" and post-cognitivism in terms of "problem defining".

5.2.3.5. Comparing Cognitivist and Enactive Conceptions of Cognition and Intelligence

Here now we can clearly juxtapose the operative definitions of cognition, as I have defined them, of each paradigm:

> Computation: rule-bound information-processing
>
> Sensemaking: open-ended, endogenous adaptive discernment of meaning by an autopoietic organism, "for, and by means of, itself"

The starting point for making sense of the difference between the cognitivist and enactive conceptions is the *already-givens* of the cognitivist account of cognition as computation. The most important already-given in the cognitivist picture is the very cognition situation: the problem, the problem-solver, and the relevant information. Computation is a description of a process and relations of these things. Because they are taken as already-given, why the problem in question is a problem, and thus why anything matters, is unaccounted for. In the cognitivist picture, computation only starts once the problem is defined and the pre-existing problem-solver knows what matters. This shows up in the orthogonality thesis when it claims that this capacity to discern what is worth solving is orthogonal to the capacity to solve a given problem. It shows up in the instrumental accounts of intelligence which define intelligence as the capacity to achieve goals, excluding the capacity to discern the goal or its significance in the first place.

From the post-cognitivist and enactivist perspective, this is only part of the story, an arc of the circle. Post-cognitivism defines the "problem" of cognition more inclusively. Determining what needs to be dealt with, what the "problem" is, is part of the problem. In short, cognitivism operates with exclusions that are taken as already-givens, and post-cognitivism includes them. In this way, the two paradigms need not be at odds with one another.

In order to compare the two accounts in a little more detail, first I will provide a reminder of the computational conception of cognition I presented in the previous chapter. Then some of the significant differences can be noted, in a way that brings into relief the solving-defining difference. I will finish with a navigational metaphor to a more intuition-friendly expression of the difference.

In the third chapter, computation was presented as "rule-defined information-processing". I spelled out this definition using the notion of Turing machines, a basic cognitivist model for mind which, arguably represents the fundamental cognitivist conceptions of the cognitive process, even if particular models

of computation, like the Connectionist, no longer strictly speak in terms of the dynamics of a Turing machine (Wheeler 2005: 104). I emphasised that the model assumes the already-given existence of everything from the particular information input, the algorithm or rule according to which the information is processed, and the problem in question, not to mention the identity of the Turing Machine itself. The significance of these assumptions is that nothing in the computational process requires discerning what the problem is or why anything is relevant. In the terms of this chapter, what is significant is defined as such to begin with by an external observer.

Again, this is the significant difference between computation and sensemaking. In the case of computation, as a model of cognition, what is relevant, significant, and meaningful to the agent is already-given in the model. The entire context and Frame of reference is already-given. From the perspective of the enactive account, it is necessary that these things are defined endogenously for them to be meaningful *for the agent*. Otherwise we get "competence without comprehension" and "Chinese Room" scenarios in which there is processing, but not sensemaking; we get AI and superintelligent AI which can solve a problem without a sense of what is relevant to us, (recall the "King Midas" problem), or even why the problem is a meaningful thing for *them* to solve in the first place. In these cases there is computation, but not sensemaking.

The Orthogonality Thesis, which we can now say is a conceptual consequence of the computational account of cognition, claims that a capacity to solve a problem and a capacity to discern whether it is a relevant problem to solve are "orthogonal axes along which possible agents can freely vary" (Bostrom 2012: 74). They are more or less independent functions which, in the "space of possible minds" need not necessarily be bound together. At this point, it should make sense that the Orthogonality Thesis is inconsistent with the enactive account of cognition in two closely related ways.

Firstly, for sensemaking, solving a problem and defining it, which is to say, discerning it as a relevant feature of the world to navigate, are parts of a cognitive process that makes up a whole circle. Solving a problem and defining it are arcs of a circle, not orthogonal. They are conceptually separable functions to an external observer who aspires to build a cognitive system, but for an autopoietic system, there is no distinction there because features of the world are always interpreted relative to the integrity of the self. More simply, orthogonality is inconsistent with a system which endogenously defines itself and, in that way, defines what is important to and for itself. What matters is not orthogonal to its capacity to achieve it. Orthogonality makes sense for systems whose definition and individuation as a whole "exists outside the watch" in the mind of observer whose frame of reference makes it relevant to build such a thing, like Shannon and his electrical communication systems.

This individuation of parts and wholes is the second, or corollary, way in which orthogonality is inconsistent with cognition as sensemaking. In the cognitivist paradigm, problem-solving and problem-defining are seen as modular and "orthogonal" functions and axes in the "space of possible minds". In a paradigm where wholes are not autopoietic, but exogenously defined and produced, the role of the external observer is, ironically, overlooked, to the effect that what in autopoietic sensemaking is a self-producing whole, is assumed to be an arbitrary assemblage of parts assumed to exist independently of that whole. Again, more simply, the second way in which orthogonality is inconsistent with sensemaking is that it assumes that the existence of cognitive "parts" precedes the whole. In cognitivism, an external observer defines a line in the cognitive process which, in autopoietic sensemaking, does not endogenously exist. The partitioning is arbitrary, relevant only to the external observer and their frame of reference.

In presupposing as already-given the nature of the agent, a computational account does not account for why those putatively pre-defined and separate functions might come as a whole. In the enactive account,

the notion of autopoiesis accounts for the self-production of the agent, which is to say, why it shows up as the particular whole that it does. The whole and the parts exist "for and by means of" one another, owing their "*presence to the agency of that other*" (Kant 2007: 202) This is also perhaps why the binding or combination problem is more of a concern for the computationalist account than the enactive (Thompson 2007: 53). It stipulates from the beginning that certain parts exist. Why those parts, and why those parts together?

Moving away from the orthogonality thesis and back to the general differences between the paradigms, the absence of a sense of relevance also speaks to the absence of a notion of anything being at *stake* in a computationalist account of cognition. Insofar as the problem for a computational system is really a problem for the external observer who designed and built the system, that system does not have a stake in the quality of its discernment. Other than in terms of expected utility, as exogenously defined, the system is indifferent to the outcome. By contrast, an autopoietic agent enacts sensemaking "for and by means of itself", coming to know by the active disruptions to its integrity it encounters as it moves. Its life is on the line. The difference makes a difference to it. In the case of the computational agent, there is the under-explored question of whether there really is a "self" there. If there is, it exists and is defined exogenously, not for "itself". It has no stake in "itself" the way that an autopoietic organism does. In this way, meaningful action is defined exogenously in a computational context and endogenously in an enactive context.

Moving onto the respective notions of intelligence, recall that in the cognitivist and computational account, intelligence is an instrumental capacity to "achieve goals in a wide variety of environments" (Legg and Hutter 2007: 9). From the enactivist perspective, those goals, and the frame of reference of the "environment" are, again, all things which must be assumed as already-given by the cognitivist. This has important fall-out vis-à-vis a conception of intelligence.

For one, intelligence as *instrumental* intelligence conceptually falls out of a conception of cognition in which discerning what is relevant is not required. If the goal is assumed, then what is rational and of expected utility *instrumentally* emerges too. Recall some quotes from previous chapters:

> For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning … Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. (Bostrom 2012: 73)

> "In short, *a rational agent acts so as to maximize expected utility.* It's hard to overstate the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines." (Russell 2019: 23) (emphasis in original)

Note the implicit stance of the external observer. The frame of reference is not that of the agent or machine, but an external agent who has a stake in the matter.

> "By "intelligence", we here mean something like skill at prediction, planning, and means-end reasoning in general. This sense of instrumental cognitive efficaciousness is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be." (Bostrom 2014: 107)

A second conceptual consequence is the way in which intelligence comes to be conceived of as a matter of optimisation.

> "The notion of an "optimization process" is predictively useful because it can be easier to understand the target of an optimization process than to understand its step-by-step dynamics." (Yudkowsky 2008: 10)

Optimisation only works in well-defined contexts. When the dimensions of a box are well-defined, a question of how to pack it in order to fit the maximum number of oranges in it, or, imagining it as a landscape, how to traverse it to an already-defined end location, are both optimisation questions.

We cannot optimize what we do not understand though. If we don't know what is relevant and meaningful, how do we know what to optimise for? Moreover, when we do optimise for something, it often comes back to bite us, as the various scenarios of AI existential risk illustrate as a metric of civilizational flourishing. It is only later that we deal with the externalities of what was not included and not defined into the problem as relevant. Historically, we are now realising that optimising for economic growth coming out of the Industrial Revolution is connected with climate change and socioeconomic inequality. In this way we assumed the definition of the problem and, much as King Midas, are dealing with the consequences of aiming for something we do not understand. Problem-solving computation admits therefore of optimisation, but sensemaking does not because optimization is not the point. At the top of this chapter, I quoted the work of James P. Carse in *Finite and Infinite Games*, in which he offers the following description of the differences between finite and infinite games. It carries insight here too for distinguishing between computation and sensemaking respectively: "There are at least two kinds of games. One could be called finite, the other infinite. A finite game is played for the purpose of winning, an infinite game for the purpose of continuing the game." (Carse 2012: 1).

A final, important note related to this broader context is that a cognitive system which does not account for its own possibility is extractive in the sense that it relies for its existence on resources that it does not generate. Kate Crawford discusses this in concrete detail in *Atlas of AI*, which will be discussed at length in the next and final chapter. In more abstract terms, computational cognition has to assume the conditions for its possibility. Sensemaking accounts for it via the notion of autopoiesis. Sensemaking is cognition that is concerned with keeping the cogniser alive. Cognition takes that for granted, and so does not include the ability to keep the cogniser and cognition going. It has to assume that the cannonball is in the cannon. In this respect, it relies on something external to keep it going. It must rely on something external for its own possibility. In this way, and to this extent, it is extractive.

The cannon is an informative example. The difference between computation and sensemaking is analogous to that between "work" and a "work-cycle" which I discussed in the section on complexity and autopoiesis – hence computation as an "arc" and sensemaking as the whole "circle". Recall that work is the expression of energy through particular degrees of freedom (Atkins 1984, Kauffman 2019: 19). Those constraints are necessary for the expression of work. Without constraints, energy dissipates in no direction at all. Insofar as the freedom-defining constraints afford and express work, a frame of reference is supposed in which it is relevant to direct energy in the direction that it does. In this way, "work", like computation, supposes what is relevant. Further, like with Kauffman's cannon, the work ends when the cannonball hits the ground, "Halting". It requires work to begin the process again.

Computation, like work, does not account for the condition of its own possibility. That comes in with a "work-cycle" which we see in the "constraint-closure" of autopoietic, complex systems. They define their own degrees of freedom for work to be expressed in continual self-generation.

To conclude this comparison, I will now consider how each might be understood in a metaphorical context of navigation, remembering that "problem-defining" is being understood in this thesis as the establishment of orientation and the discernment of meaning on that basis. So, we can ask, what does navigation look like to computation, and how do we make sense of navigation with sensemaking? What follows is speculative and metaphorical attempt to make sense of the difference between these two conceptions of cognition. I am appealing to human experience of navigation in the broadest sense of the term.

As has now been over-laboured, computation assumes the definition of the problem and is concerned with solving it. In the context of navigation, this might amount to having a well-defined map of the terrain and determined locations of departure and arrival. The "problem" is already defined, it is a matter of solving it. A map is not a necessary feature though, as Brooks' sensorimotor approach ("the world is its own best model") shows. There will be many ways to get to the objective. The instrumental rationality and intelligence of computation involves optimising a route to the destination given the pre-defined variables, i.e. whether it is relevant to get there quickly, safely, along a "scenic route" and so on. Perhaps the weirdest feature is that "success" and "performance" at the navigation is determined by an external observer.

Navigation from the perspective of sensemaking is radically different. There is, primarily, no map, but there is also nowhere to go in the sense that no final goal is assumed. The game is "infinite" and about keeping the game going (Carse 2012). What is meaningful is endogenously discerned relative to the matter of staying alive. Leveraging the metaphor, we might say that the landscape is unknown and, perhaps, clouded in fog, requiring a capacity to discern features, discern location, what is going on and what, if anything, is the "problem".

In French, the web of meaning encompassed in the etymological sibling of "sense" – "sens" – is illuminating here. Among the many senses of the term "Sens", it can simply mean the same as in English – "sense" - both in the sense of the five senses, and in the sense meant in this very performance, namely, meaning, in some sense. It can also mean "direction" or "way", in the sense of giving directions: "dans quel sens?" ("which way" or "in which direction") – "that way"; "in that direction". The meaning of "sense" and "Sens" therefore distinctly entangles the notion of *meaning* and *orientation*. Meaning, orientation – *navigation* - are each and all interwoven aspects of sense-making, when understood in this sense[81].

Indeed, sense-making as "orientation" suggests that we are situated somewhere, trying to orient ourselves given what can be seen from our perspective. *If we could see the whole territory at once, navigation would be a non-cept, not a concept or thing, for the issue or "problem" would never arise.* And, of course, not all organisms move, so the navigational metaphor may be unhelpful for making sense of the sensemaking of plants, but plants are still autopoietic, in need of adaptive discernment, enacted "for, and by means of, the self", the autopoietic whole.

[81] A worthwhile question here is whether this kind of thinking is applicable to organisms which do not move, like trees and other plant life. There is considerable literature on "plant cognition", like (Mancuso 2018), and not space here to adequately discuss the differences. Suffice it to say that whilst plants may not to the extent that animals do, plant life is nonetheless embodied and can be said to engage in sensorimotor sensemaking in this regard, in a manner unique to its morphology (Sheldrake 2020, Di Paolo et al. 2017).

The point of this wordplay and speculation is to connect sensemaking with problem-defining and the orientation involved in establishing a basis for discerning meaning.

Having now discussed life-mind continuity, autopoiesis, and sensemaking, we can now turn to the last of the four concepts of the enactive theory of cognition.

5.2.4. Enactive Dynamic Co-Emergence of Self and World

Enactive dynamic co-emergence of self and world is the fourth and final concept with which I am endeavouring to make sense of the post-cognitivist paradigm of mind in terms of "problem defining". Something which both cognitivist and post-cognitivist ideas share is that cognition is an interaction across the individuating threshold between an agent its environment. The basic idea of "enactive dynamic co-emergence of self and world" is that neither autopoietic self, nor environment – neither "side"– precedes the other. Rather than they dynamically co-emerge together. In this way, following what might well be a slogan of enactivism, *being brings forth a domain of significance* (Thompson 2007: 58-60).

Another way I have pointed in this direction in this thesis is the idea that particular conditions of observation generate particular observations. The takeaway from this section concerns the consequences of this insight. Different beings have or enact different conditions of observation and so observe different things. There is something of a paradox here insofar as it suggests that "the world" is different relative to each agent capable of awareness and perception. It suggests that we, as different individuals, both are and are not navigating the same world. The case is open on these ontological questions. What is relevant is to supplement the "problem-defining" character of post-cognitivism with the caveat that we can only "define" the world in terms relevant to us, much as our hands only hold what fits their grasp.

The key idea is that, as an autopoietic being produces itself, its particular range of sensory apparatuses are stimulated as it moves about the world. These stimuli articulate the features of its world. *Its world is the domain of features to which it is sensitive*. In this way, as an autopoietic being produces itself, it also produces its world. Much as when a robot is endowed with particular sensorimotor faculties, it can only navigate relative to those features. An agent with "measurement" capacities, can only "perceive" the world as per those sensibilities and capacities of measurement. While few would say that the bounds of its perceptions mark the bounds of *the* world, or what exists independently of it, the bounds of its perception do mark the bounds of *its* world. In this respect, things are real for a being in the way in which, and to the extent that, they endogenously make a difference to it.

So, it is *not* the case, according to the enactive theory, that the autopoietic being and subject causally produces the object and the "world". Again, it is not the case that one precedes the other. They dynamically co-emerge. It is helpful to describe the process of autopoiesis in this subject-object way, the problem is that it suggests a causal locus and genesis in the agent. Strictly speaking it is probably more appropriate to remove the subject and say something like "there arises autopoiesis producing itself", in the same way that we might say of Escher's *Drawing Hands* (fig. 1) that "there arise hands drawing themselves". There is a strong association in the enactive literature with Buddhist contemplative insight with respect to this knot and where to place the agency (Wheeler 2005, Thompson 2007, Varela et al. 1991). I am labouring the point because it dominates intuitions to think that either environment precedes agent, as per the objectivism of cognitivism, or to misinterpret autopoietic sensemaking as producing the environment.

Varela et al. (1991: 172) talk of this as a chicken-and-egg problem, articulating the "chicken position" as the sense that "the world out there has pre-given properties" and the "chicken position" as "the cognitive system projects its own world, and the apparent reality of this world is merely a reflection of internal laws of the system". "Dynamic co-emergence" is a middle path of sorts, though, it is more precise to say it is a transcendental description of the condition of possibility of an agent's cognition of its world. There is a transcendental fit between what an agent is sensitive to, which is to say, what it is capable of cognising, and what it perceives. Like the example of a hand grasping only what fits, dynamic co-emergence means that the world for an organism – their world – is the world they are capable of encountering and perceiving.

So, to say the enactive individuals "bring forth their own environment" or that "being defines a domain of significance" does not mean that an individual enacting itself causally brings the world and its contents into existence. The dynamic is not one of linear causal sequence, one thing existing and then causally bringing about another. What is being said is not something like "my existence causally brings into existence the moon and the world". Again, neither self nor world has precedence over the other. Rather, it is a relationship between *my* existence and that of *my* world. As I am produced, so is the world of features to which, given my own production, I am sensitive. In this way, Thompson, citing Merleau-Ponty notes: "the environment emerges from the world through the being or actualisation of the organism." In the case of animal life, the environment emerges as a sensorimotor world through the actualisation of the organism as a sensorimotor being." (2007: 59)

Thompson later summarises, claiming that "*Emergence of a self entails emergence of a world.* The emergence of a self is also by necessity the co-emergence of a domain of interactions proper to that self, an environment or *Umwelt*. …**The organism's environment is the sense it makes of the world**" (Thompson 2007: 158) (italics in original, bold my own)

This last line speaks to what in philosophy is sometimes referred to as the Euthyphro question from Plato's *Euthyphro* dialogue (Plato 2009). In the dialogue, the question is posed, is the pious loved by the gods because it is pious, or is it pious because the gods love it? The question can be generalised to "does the world make sense or do we make sense of it?" Put this way, the phenomenon of dynamic co-emergence is a response to this same question, but one which slips the either-or economy of answers with its transcendental move: dynamic co-emergence is the transcendental fit between what an organism perceives and what it is capable of perceiving, or, in terms of sensemaking, it is the transcendental fit between the sense an organism makes, and the sense it is capable of making[82].

The world unique to an organism is spoken of in many places in terms of "[Jakob] von Uexküll's original notion of an Umwelt. An Umwelt is an animal's environment in the sense of its lived, phenomenal world, *the world as it presents itself to that animal thanks to its sensorimotor repertoire.*" (Thompson 2007: 59)

> "An environment, in von Uexküll's (1957) sense of an *Umwelt*, has meaning and value…An organism's environment is not equivalent to the world seen simply through the lens of physics and chemistry. Physical and chemical phenomena, in and of themselves, have no particular significance or meaning; they are not "for" anyone. Living beings shape

[82] What this means for the Euthyphro question specifically is trickier than this because it goes beyond "mere" cognition of the world to a sensitivity to the divine. As such, the transcendental move corresponding to the notion of enactive dynamic co-emergence would involve claims about the divine. I am unaware of in-depth work discussing the Euthyphro problem from the perspective of Enaction. It is unclear to me from here what the implications of such work would be, but it seems like it could be worth exploring.

> the world into meaningful domains of interaction and thereby bring forth their own environments of significance and valence." (Thompson 2007: 153-154)

Some concrete examples may be helpful. I cannot see in UV or infrared, as some animals can. This does not mean that they create that bandwidth of light or what they perceive in that medium. My world, my "umwelt" is defined by what I am able to see, what I am capable of seeing. All sorts of things will make a difference, from the physiology of my optics and nervous system, to both my personal history and the history of the physical and memetic environment in which I am embedded, to my present state, physiological and psychological.

Consider what an experienced rock-climber sees when she looks at the Dawn Wall in Yosemite Park compared to what I, an unexperienced climber, sees. I see nothing, no possible means of scaling the wall, only wrinkles and dimples in the surface of the wall which suggest it to be an absurd undertaking. For the experienced climber, these features, and many others I will not notice, show up as genuine points of purchase and leverage on the wall, affording a climb. There exist features for her, but not for me. We would not say that she brings those features into objective existence, only that, in the course of her development of her capacities of perception and discernment – "training" or "education" – she became sensitive to those features. In the language of enaction, they emerged in the course of her autopoietic actualisation, they were "brought forth".

In this context, questions of mind-dependence or independence and objective or subjective realities is not so much the interesting one. Something exists for both of us. She, "being a frame", makes sense of the wall according to those features which show up for her as a particular "condition of observation", as I do according to the frame and condition of observation that I am. Being "experienced" in this context, the "training" she has undertaken, amounts to defining constraints through which she expresses her embodied agency, and "working" through those constraints again and again, becoming more and more experienced at operating through those constraints on her agency and body. The constraints in question are constraints which she defined, insofar as she continually chose to pursue the activity. She defined the constraints which in turn constrained her to "work" through particular degrees of freedom, generating, in time, an ever more fine-grained coupling between self and environment. In one sense she does not create the wall, but in another she does because whilst the wall may be full of features for her, for me, the wall shows up as a mostly featureless tower of granite. My coupling to the Dawn Wall is not so fined-grained as hers. For her, it is rich with information, affordances, constraints and work, potential and possibility. In this way, she sees what she sees because of what she is, hence the enactivist credo that "Being defines a domain of relevance."

Continuing, we can say then that a condition of observation defines a possibility or domain of certain observations, regardless of what is there independent of that condition of observation. When I am not looking through a telescope or microscope, I do not see what is possible to see through them, but of course I do not deny the existence of any of what is observable through those lenses. I recognise that as an autopoietic condition of observation, I perceive certain things, and when my condition of observation is changed via anything that might make a difference – tools, psychoactive plants and substances, culture, gender, race, love - I perceive differently. I need not imagine that that which was perceived in another condition now necessarily does not exist. I enactively perceive *my* world. In terms of

sensemaking, I am not concerned with *the* world. I am concerned with what is real *for me*, meaningful to me. That is the "world" enactively "brought forth"[83].

Now we have come to the end of the discussion of the enactive theory of mind and cognition. This could mark a sufficient account for getting a sense of the way in which post-cognitivism can be characterised in terms of problem-solving, but because enaction is itself part of a larger whole, the other "E's" of 4E cognition, it is worthwhile for a characterisation of post-cognitivism to show how the other E's too can be understood in terms of problem-defining.

It is important for another reason too. As will be discussed, as far as their plausibility as contributions to a theory of mind is concerned, the notions of embodiment, extendedness, and embeddedness are predominately so-far evaluated in terms of how they contribute to *solving* problems. So, though these notions are already departures from cognitivist thinking, they are being captured by a problem-solving kind of thinking[84].

5.3. The Other Es: 4E Cognition

In this subsection a very basic overview of 4E cognition will be given. The classical 4E's of cognition are embodiment, embeddedness, extendedness, and enaction (Newen et al. 2018). The discussion leading up to this was on enaction, covering life-mind continuity, autopoiesis, sensemaking, and dynamic co-emergence, so the discussion from here will be on the other Es. Work in this area is primarily conceptual, as opposed to being a practice of cognitive science, as it is establishing the theoretical foundations of a new paradigm of mind research. Here is an introduction from a recent textbook on 4E cognition:

> According to proponents of 4E cognition, however, the cognitive phenomena that are studied by modern cognitive science, such as spatial navigation, action, perception, and understanding other's emotions, are in some sense all dependent on the morphological, biological, and physiological details of an agent's body, an appropriately structured natural, technological, or social environment, and the agent's active and embodied interaction with this environment. (Newen et al. 2018: 3)

As a model of cognition exploring these phenomena, 4E cognition shares with other models the basic question of inquiry "what is cognition?" Traditionally, this "what" question has been more or less universally engaged via the heuristic of a "where" question. That is, the "what is cognition" is often engaged with the simpler, "where is cognition" question. This "where" reveals the basic asymmetry of the cognitive situation which, as a whole, is a system composed of two coupled systems, systems we

[83] There is a danger of a certain solipsism here: if I am enacting and bringing forth *my* world, how do I interact with yours or anyone else's, and how do we create the consensus reality of "our" world? Put in other terms I have used so far, how do we put together conditions of observation and their observations - "my" world, "yours", and "the" world? There is existing work exploring the domain of the social (and cultural) in ways that pull on the principles of enactive theory and post-cognitivism more broadly. See (Di Paolo and De Jaegher 2022, Di Paolo et al. 2010, Barad 2007).

[84] I have a paper in preparation on this topic: "Making Sense of 4E Cognition: Problem-Solving and Problem-Defining".

usually refer to as "agent" and "environment", self" and "world", "subject" and "object", "knower" and "known", "observer" and "observed" (Maturana 1987, Bateson 1979). The asymmetry is sometimes understood in terms of "agency", whereby the subject has the agency, and the object does not. It is also understood in terms of information-flow (Krakauer et al. 2020) – recall the notion of "symmetry-breaking" in the discussion from complexity science. Krakauer et al. use this asymmetry to individuate entities, both individuals and collectives. As Karen Barad articulates it, "knowing is a matter of part of the world making itself intelligible to another part" (Barad 2007: 184). Perhaps the fundamental point then is this partition, with cognition being a particular kind of activity across the partition, in asymmetric fashion such that we would locate agency on one side more so than the other. Given this individuation of one part of the world from another, there is some asymmetric interaction across the threshold, be it in the currency of energy, matter, or information. Wheeler speaks of an "ontological distribution" (2010: 2).

In this light, the difference between computational accounts in cognitivism and enactivist accounts in post-cognitivism is that the computational model assumes this individuation of self and world – the cognitive situation – as a precondition for talking about cognition, and the enactivist model accounts for how that individuating threshold is generated with its notions of autopoiesis and dynamic co-emergence. In both cases cognition is some sort of asymmetric interaction across that threshold, but so far only the enactivist account speaks to how it is initially generated.

So, the "what" of "what is cognition" is usually engaged in terms of the spatial distribution (aka "where") of this dynamic, seeking to identify some sort of high-density locus "where" cognition is located. One way of understanding the significance of the 4E's, therefore, is how radically it rethinks "where" cognition is "located" – it is no longer located only in the brain but extends out of the head to include tools and environment.

It is important to mark the different ways the Es are understood because whilst "essentially every part of the cognitivist view has been challenged in the 4E literature" (Aizawa 2018: 3), it is still in principle possible to understand each and all of the 4Es in the language of computation and with the intuitions of cognition as "problem-solving". See for example work by Brito and Marques (2016), Villalobos and Dewhurst (2017, 2018), Kersten et al. (2017) for discussions at this juncture. That said, whilst it is in principle possible to articulate the ideas of 4E cognition in these terms, in practice 4E work doesn't, and explores its own way of framing things.

However, from the perspective of this thesis it is not clear that, at least in existing discussions, the 4E's really do challenge every part of the cognitivist view as Aizawa claims. Of course, in my presentation here of the two paradigms so far, there is indeed very little overlap. The concepts are different, and, as I have emphasised, where each model draws the boundary of the cognitive process is also different. However, as long as 4E conceives of each of the Es in terms of how they each contribute to how an individual solves a problem, we are still viewing it through a cognitivist lens.

To the extent that this is the case, 4E cognition is *not*, then, in its current articulations, a major departure from computationalist views. Anchored in the intuition that cognition is still some sort of problem-solving, just now not exclusively restricted to the head, the dynamics of 4E cognition become more or less amenable to description in computationalist terms and all the baggage that comes with it[85]. As I

[85] Some authors of the enactive persuasion may deny this on the grounds that computations involve representations, and enactivism, so the thinking goes, is a distinctly non-, or even anti-representationalist theory of cognition. I claimed in the cognitivism chapter that the question of representation is derivative, having been introduced decades after computation was already established as a technical notion for cognition, and that, at least

laboured to show in the cognitivism chapter, computation and a problem-solving view of cognition fit together such that, if we start with one, we can be led to the other.

Therefore, in order to appreciate the significance of the 4E's, from the perspective of this thesis I want to demonstrate the 4Es in a way which adequately captures their novelty in relation to the cognitivist model. The best way to do that I think is to understand each of the 4Es as specifying *conditions of cognition*. This is to say, they are the features of the cognitive situation which cognitivism takes as already-given. Recall that autopoiesis enacts or "brings forth" a domain of significance, defining a domain and world of significance for an organism, the condition of its cognition.

In this line of thinking, we can for example understand the significance of embodiment in the following way. Our body defines certain constraints and degrees of freedom on the expression of our choice, energy, agency, and more specifically sensorimotor cognition. I cannot physically touch and make sense of what is beyond my physical reach. The significance of Extendedness, from the problem-defining perspective, is how a tool changes or effects the constraints of my cognitive situation[86]: among other things, in a rainstorm, a waterproof jacket changes the demands on me, my "problem" is different than without it. Similarly, "embeddedness" can be understood as the (eco-logic of) environmental constraints in which the agent exists, and thus which define the situation in which the agent exists.

In any case, the distinction between problem-solving and problem-defining is not something that is employed in the literature. Because the Es are often understood in problem-solving ways, in the coming sections, I present both and compare both understandings of each "E". Because enactivism was the focus of the chapter so far, I focus, in what follows mostly on the other E's, with a summary reminder of enactivism at the end.

There does not as yet seem to be coherence or consensus within the 4E approach as to how to understand each of the E's or their broader significance vis-à-vis mind and cognition. The approach of this thesis is that, in order to properly distinguish the 4E's *in kind* from an implicitly computational and reading in which they become just another theory about how the mind solves problems, it is important to first understand how the 4E's define the conditions and generative constraints of mind and cognition. Otherwise, the fundamental story is the same as that of the Turing Machine which forms the basis of cognitivism.

5.3.1. Embodiment

In the chapter "Building a stronger concept of embodiment" in the 2018 volume *The Oxford Handbook of 4E Cognition*, Gallagher readily notes that:

> "As we see in the present volume, however, there is no consensus theory of embodied cognition (EC), and debates continue about the best way to understand this notion. The

for my purposes, the distinction between problem-solving and problem-defining better distinguishes the paradigms. I stand by this.

[86] This echoes Marshall McLuhan's work and famous line that "the medium is the message" (McLuhan 2005). I mention it again in the section on extendedness. Work connecting the extendedness of cognition with McLuhan's work in Media Studies on "the extensions of man" is worth further exploration because McLuhan articulates a radically detailed "problem-defining" view which has much to offer in the context of cognition and may help us avoid reinventing the wheel.

> alternatives range from conservative models that remain close to cognitivist conceptions of the mind, to more moderate and radical camps that argue we need to rethink our basic assumptions about the way the brain and the mind work. Most recent debates have been focused on the pragmatic and action-oriented perspectives of ecological, enactive, and extended conceptions, which either minimize reliance on the notion of representation or eschew it altogether." (Gallagher 2017: 1-2)

For a notion which all but identifies a distinct paradigm of mind research, a tractable, accepted definition of embodiment is difficult to find. Echoing Gallagher, in three different reference works including the above volume (Newen et al. 2018) on 4E cognition, *The Routledge Handbook of Embodied Cognition* (Shapiro 2014), the Stanford Encyclopaedia Page on embodied cognition (Shapiro and Spaulding 2021), a settled definition of embodiment is notable for its absence. Instead, a categorisation is made of three different ways of understanding what embodiment might mean. These conceptions will be specified momentarily, but first it is worth noting some interesting reasons why a clear definition might be elusive.

In the cognitivist paradigm, there does seem to be central ground on which different theorists stand together in defining computation. Perhaps because 4E cognition is in a more nascent stage of development, there is not yet such basic consensus beyond a trivial sense that the body and mind have some kind of metaphysical relationship. This comes out in *The Embodied Mind* (Varela et al. 1991) in which a clear definition of embodiment is not yet forthcoming. Being one of the earliest book-length works on the matter, the aim of the book was arguably more to articulate a new path of mind research focused on integrated cognitive science with experience, and less about nailing down a particular theory of embodiment. That said, the heritage it recognises in the phenomenology of Merleau-Ponty (2002) to thinking about embodiment, offers a big picture at least of the significance of embodiment for the new direction of research. And one thing the phenomenological reading of embodiment did bring into relief was the way "embodiment" had otherwise theretofore been concerned simply with the *behaviour* of physical systems, biological and otherwise.

> "We hold with Merleau-Ponty that Western Scientific culture requires that we see our bodies both as physical structures and as lived, experiential structures – in short, as both "outer" and "inner", biological and phenomenological. These two sides of embodiment are obviously not opposed. Instead, we continuously circulate back and forth between them. Merleau-Ponty recognised that we cannot understand this circulation without a detailed investigation of its fundamental axis, namely, the embodiment of knowledge, cognition, and experience. For Merleau-Ponty, as for us, *embodiment* has this double sense: it encompasses both the body as a lived, experiential structure and the body as the context or milieu of cognitive mechanisms." (Varela et al 1991: lxi-lxii) (italics in original)

One place there is a more explicit description of embodiment is in Evan Thompson's (co-author of the above work), (2007) book *Mind in Life*:

> "The central idea of the embodied approach is that cognition is the exercise of skilful know-how in situated and embodied action (Varela et al. 1991). Cognitive structures and

> processes emerge from recurrent sensorimotor patterns that govern perception and action in autonomous and situated agents. *Cognition as skilful know-how is not reducible to prespecified problem solving, because the cognitive system both poses the problems and specifies what actions need to be taken for their solution.*" (Thompson 2007: 11) (emphasis my own)

Thompson's description here more or less explicitly defines the skilful know-how of embodied cognition in terms of problem-defining when he says that the cognitive system "poses the problems and specifies what actions need to be taken for their solution". However, such an understanding of embodiment does not seem to be common.

Returning to the reference works above, Shapiro and Spaulding (2021) note three different ways of understanding the significance of embodiment: conceptualisation, replacement, and constitution.

In reverse order, understanding embodiment in terms of constitution sees the body not (only) contributing causally to cognition, but as potentially constitutive of cognitive processes. Here already there are both cognitivist and post-cognitivist interpretations. Generally speaking, a cognitivist would be happy to say that a body plays a constitutive role, in the same way that a substrate of some sort is necessary to realise the algorithmic computations. In this interpretation of "embodiment", it functions merely as a means of physical implementation or realisation of the computations necessary to execute an already-given function. That said, there exist analytic nuances in the form of distinctions between "strong" and "weak" embodiment, and whether the "bodily" constitution of the cognitive process is realised by bodily or "extrabodily" processes:

> a. A cognitive process is *strongly embodied by bodily processes* if it is partially constituted by … processes in the body that are not in the brain;
>
> b. A cognitive process is *strongly embodied by extrabodily processes* if it is partially constituted by extrabodily processes;
>
> c. A cognitive process is *weakly embodied by bodily processes* if it is not partially constituted by but only partially dependent upon extracranial processes (bodily processes outside of the brain);
>
> d. A cognitive process is *weakly embodied by extrabodily processes* if it is not partially constituted by but only partially dependent upon extrabodily processes. (Newen et al 2018a: 4) (italics in original)

These analytic formulations strike as a bit confusing, particularly where it talks about a cognitive process being constitutively embodied by something outside the body (b) and (d). This is how "extendedness" is understood, and (d) in particular is "identical with the property of being embedded" so the four E's overlap (Newen et al. 2018a: 4). This overlap is to do with different positions people take on whether "bodily processes" play a causal or constitutive role, and the way in which "extrabodily" processes count as participating in the dynamic. In the following sections of extendedness and embeddedness the distinction between causal and constitutive will become clearer. The above four formulations aim to overview and include the idiosyncrasies of a large span of 4E literature so are perhaps more "complete" than they are helpful in making sense of what "embodiment" means.

Whilst a metaphysically developed theory of embodiment will want to make such nuances, the fundamental significance of embodiment is the departure in a model of cognition from a sense that all the relevant work is in the "software", and that "hardware" is merely a means of physically realising the software. The role of the body moves from being an uninvolved substrate to participating in some sense, whether causally or constitutively. That said, it is not always clear what the difference is between a body actively "constituting" a cognitive process (a 4E conception) and a body being an inert substrate or medium for "realising" it (a cognitivist conception).

In support of the ideas of this thesis, it seems as if this blur could be to do with the problem-solving orientation with which even notions of embodiment are being explored. If a body is to be understood in the context of cognition in terms of how it helps solve a problem, then it makes sense to imagine that role in causal or constitutive terms, which leads us to a conception of embodiment that is fundamentally not much different from a "substrate" conception. "Causal", "constitutive", and "implementational" are all fairly functional roles. It is the position of this thesis that this lack of real differentiation from a cognitivist picture is because embodiment is still being explored, however implicitly, in terms of how a body (functionally) contributes to *solving* a problem.

A notable example here is the paper by Müller and Hoffman (2017) in which they explore the notion of "morphological computation". They distinguish cases in which morphology "facilitates" control, enabling "physical behaviour in the real world" (ibid: 16) or "facilitates" perception, endowing sensory perception of a certain kind (Müller and Hoffman consider the vision of a fly and a mouse) or "morphological computation proper" in which the body operates as a computational substrate, actually performing computations. Even without the language of computation, in all cases the question is still how the body in some way facilitates solving some problem, so is still a cognitivist interpretation as per the problem-solving characterisation I am using in this thesis.

So, that is the constitutive sense of embodiment. Moving onto "replacement", replacement understands embodiment to be among concepts which can *replace* more computationally inspired notions of cognition like symbols, representations, etc. (see e.g. Chemero 2011). Here the concern seems to be discerning the concepts which enable a conception of cognition to come closer to the truth.

Finally, "conceptualisation" is the reading of embodiment perhaps closest to a problem-defining reading of embodiment. Inspired by the work of Lakoff (1998), Lakoff and Núñez (2000), and Lakoff and Johnson (2008), Shapiro (2019) defines "conceptualisation" in terms of the heavy "body-dependence" of our concepts and metaphors. Our situation in space means that all motor actions are asymmetrically coordinated in space such that we have a basic set of concepts like "up" and "down", "left" and "right". It is close to a problem-defining account, but even if it is granted that a body defines the space of possible concepts, it seems odd to talk about the embodied mind in terms of abstractions like concepts. "Conceptualisation" says nothing of how such a body defines or navigates a *physical* world with *physical* features. To this extent, it seems that the underlying conception of cognition ultimately has it as an abstract thing, which seems to put us back in the cognitivist position of thinking that cognition is about the "software".

Whilst an understanding of embodiment in terms of how it affords certain concepts might minimally count as a problem-defining take on embodiment – "the body defines the domain of available concepts" - other categories of discussion of embodiment do not seem to define embodiment this way.

Compared to these approaches, here is a "problem-defining" reading of embodiment, one still consistent with the above.

Embodiment "situates" the embodied mind in physical space. A body defines the degrees of freedom through which sense can be made, and choice, agency, and physical energy and "work" expressed. In biology, the morphology of an organism is sometimes referred to as a "Bauplan" (Willmore 2012). It translates on the one hand as "architectural drawing", and more directly or literally as "body plan", but German speakers will recognise that "bau", from "bauen", appears in a network of words in which it can variously mean things like "build", "make", "construct", or, for purposes here, "*work*". A body as "work-plan" defines the degrees of freedom through which work can be expressed. Harking back to the discussion on "constraint closure" from the previous chapter, a body enacts the generative constraints which force energy through complex degrees of freedom which, for autopoietic bodies, produce that very body.

In the way that embodiment localises an agent in physical space, it also defines a certain possibility space for an agent and their awareness, constraining the domain of possible action and interaction, its choice-space and, relative to the capacities of the body, the "landscape" of that choice space. Relative to the capacities of a body, certain choices and actions will be harder or easier. A body is the difference between "here" and "there", "situating" and locating an agent in particular place with a particular frame of reference. In this way, and to this extent, a body is a condition of observation in which certain things show up as relevant and meaningful, relative to the needs of the embodied individual.

When embodiment is enactive, autopoietically producing itself, then there is again the "dynamic co-emergence" of body and world. This means that a body enacts a world for itself, perceiving that which is sensible and relevant (for a reminder, see the discussion of autopoiesis in the previous chapter). In this way, a body not only defines, situates and locates a mind, but enactively "brings forth" a world of significance for itself. With embodiment, both "self" and "world" are defined.

Compare such an account of embodiment with one guided in inquiry by how the body contributes to problem-solving cognition. Understanding how our bodies situate us, or *are, and enact,* our situation, speaks to how and why things show up as relevant for us. With such an understanding of embodiment, the body is not there to simply facilitate, constitute, or otherwise contribute to computation to an arbitrary end. With this first "E", understood in a way that makes sense of how our bodies contribute to defining what is meaningful to us, we can return to consider what aspects of cognition are available for technology to extend. Computation is not the only thing going on, and yet, de facto, it is primarily what we speak of when we speak of technology extending our minds beyond our heads.

Turning now to this notion of extendedness as the next "E", the relationship between theory of mind and understanding of technology is very much centre stage.

5.3.2. Extendedness

Extendedness is where the "where" question of cognition really reveals itself. It shares with embodiment the basic idea again that cognition is not located only in the head, but goes further, saying that even the boundaries of the body are not the whole story. Acknowledging the various ways in which we use tools, extendedness means that boundaries extend beyond our bodies to include the objects whose capacities we leverage. In this way, tools are included in the boundary of the larger cognitive system which "solves" a problem.

Alongside problem-solving, the major features of cognitivist thinking are readily apparent in standard discussions of extendedness – the presumed identity of cogniser and their domain of significance and

goals, a fundamental functionalism, and a computationalist account of it all. I go through each before finishing with a more post-cognitivist interpretation of extendedness and connecting it back to the argument of this section.

The cognitivism (implicitly) invoked in conversations about extendedness seems to begin with the "where" question. An early and explicitly computational inquiry into the conceptual plausibility of extendedness was made via a notion of "wide computationalism" (Wilson 1994), in which the computations necessary to solve an individual's goal were considered to be performed by a system which did not entirely supervene on the internal states of that individual:

> The states (and the processes which are the transitions between such states) over which a computational psychology quantifies need not be individualistic because the cognitive system to which they belong could be part of a wide computational system. That is, the corresponding computational system could transcend the boundary of the individual and include parts of that individual's environment. If this were so, then the computational states of such a cognitive system would not supervene on the intrinsic, physical states of the individual; likewise, the resulting computational psychology would involve essential reference to the environment beyond the individual. The states and processes of a wide computational system are not taxonomized individualistically. (Wilson 1994: 354)

Note that the identity of the individual is assumed, something which in this thesis is "paradigmatically cognitivist". This is something which shows up throughout.

The "where" question about computation and cognition continues to show up too. Building on Andy Clark's (1997) *Being There: Putting Brain, Body, and World Together Again*, the "where" question is explicit in David Chalmers' and Andy Clark's seminal paper, "The extended mind". Note the opening line: "Where does the mind stop and the rest of the world begin?" (Clark and Chalmer 1998: 7).

The problem-solving orientation is also there. A few paragraphs into their paper they invite the reader to "consider three cases of human problem-solving…" in which, in each case, the solving of the problem is variously assisted by some sort of technology (ibid). The problem-solving orientation shows up in more contemporary overview discussions too: "Extended cognition in its most general form occurs when internal and external resources become fluently tuned and deeply integrated in such a way as to enable a cognitive agent to solve problems and accomplish their projects, goals, and interests." (Kiverstein et al. 2013).

An example from (Kiverstein 2018) involves a bartender who sets out on the bar the glasses for the drinks they're making to make it easier to remember which drinks they have to make as they go along. Again, the problem-solving language is present, and so are several other cognitivist concepts – information and function in particular. The problem-solving language is not a problem of course. The point of identifying it is to identify the kind of thinking that is happening. Extendedness is clearly being thought about in terms of how they contribute to *solving* problems, intentionally or not. The problem is that this kind of thinking misses how the Es of cognition contribute to *defining* the condition of cognition and "bringing forth" problems, both in the McLuhanesque sense of defining a domain of

interaction[87], and, similarly, in the way they define degrees of freedom through which agency, choice, and "work" can be expressed. So, the problem is not that the bartender example is not adequately explained – the explanation is fine – it is the kind of thinking used to make sense of extendedness. Here is the example:

> Instead of needing to store all of this information and keep it in mind, some of the work of remembering is offloaded onto the environment in the line of glasses. The environment now functions as an external store of information and performs the role of a stand-in for the drinks order. To work out which drink to serve next, the bartender need only look and reach for the next glass in the line. The information-bearing load on his working memory is thereby significantly lightened. Part of this work is delegated to the representational structure temporarily assembled in the world, which can then be used to control and guide action so as to bring the task at hand to successful completion. (Kiverstein 2018: 21)

Even ignoring the cognitivist terminology of "information" and "representation", the very nature of the example lends itself to a problem-solving sense of extendedness. The problem is already-given, so is the problem-solver. The entire cognitive situation is already-given. All that needs doing is to solve the problem. Again, this is all fine. The "problem" then is that we don't even begin to consider the possibility that our extensions might define our "problems" too. What this might look like is considered later in this section. For now the point is just to highlight the cognitivist, problem-solving way we are thinking about extendedness.

The predominantly functionalist orientation of current discussions of extendedness is also important to highlight in this regard. Sprevak makes an early and particularly clear-cut case for the relationship between functionalism and extendedness saying no less than that "functionalism entails HEC" (hypothesis of extended cognition) (Sprevak 2009: §4). In his (2010) paper, Wheeler also defends an "extended functionalism", based on a functionalism which he claims to already be fairly standard in contemporary thinking on the matter:

> "The claim that ExC [extended cognition hypothesis] is in some way a form of, dependent on, entailed by, or at least commonly played out in terms of, functionalism is now pretty much part of the received view of things (see, e.g., Adams and Aizawa 2008; Clark and Chalmers 1998; Clark 2005, 2008…)." (Wheeler 2010: 244)

Not everyone thinks about extendedness in explicitly functional terms though. Kiverstein (2018) points out that what is going on in these discussions about where the boundary of the cognitive agent is, is the matter of determining what really counts as "the mark of the cognitive" – its "properties" – and functionalism is only one way of doing that. A functionalist account of things means that tools beyond the skull which functionally participate in the cognitive process "count". Kiverstein compares this

[87] Again, McLuhan's famous idea was that "the medium is the message" (McLuhan 2005). The usual interpretation of this line is that, if we are concerned with understanding the effect of a technology, social media for example, the important effect is not the content that is presented on the platform, but rather it is found in understanding how the technology as a medium reshapes – redefines – our possibilities for action and interaction. Our extensions define and redefine our situation.

functionalist account of extendedness with theories about embeddedness (see next subsection) which agree that the environment, tools, and other things "outside" the skull participate in some sense in cognition, but which maintain that cognition still occurs more or less in the head because that is where representation, a necessary "mark of the [computational] mental", takes places (Kiverstein 2018: 18-19).

The distinctions become quite nuanced. The disagreement between different positions concerns "the cognitive status of external information-bearing structures and the bodily actions that are performed on those structures" (ibid: 21). Some think these external information-bearing objects have a constitutive role, and others think they have a causal role. What Kiverstein refers to as the "embeddedness" position says that these extensions only have a *causal* role, causally helping an agent arrive at a solution, but not *constituting* the cognitive process. The distinction is made on the basis that some kind of internal representation is necessary for it to count and constitute a "mental" process. So, because the bartender's glasses aren't computing representations, but only causally facilitating, as an "outsourcing" or "offloading", the computation of representations remains in the brain[88].

Whilst for Kiverstein functionalism in this context is reserved for only those cases in which external information-bearing structures are *constitutive* of the mental process, for the purposes of this thesis, both positions of which he speaks are functionalist in a paradigmatically cognitivist sense. This is because both positions speak of the external thing in a way that sees it functionally contribute, whether causally or constitutively, to the cognitive process of solving some problem. Kiverstein acknowledges that whatever disagreement there may be between these two ways of thinking about cognitive extensions, both positions agree that for *explanatory* purposes, reference to external extensions have to be made:

> Embedded and extended theorists therefore agree that internal cognitive processes will often not be sufficient for explaining cognitive behaviors. Given the extent of this agreement about how to go about explaining many of our problem-solving behaviors, one might be forgiven for wondering what is really at stake in this debate. (Kiverstein 2018: 22)

For the purposes of this thesis, this shared explanatory functionalism is important because it identifies thinking that is fundamentally of a (cognitivist) kind. For Kiverstein (2018) and Sprevak (2010), what matters is a metaphysical question about what counts as a mental process or not, and so the distinctions between the causation and constitution matter. As far as this thesis is concerned, they are both cognitivist positions, and so are not particularly different. This is clear from the implicit computationalism, the problem-solving orientation, along with the functionalism, and perhaps most notably, that the individual and the "external information-bearing structures" are presumed as already-given and well-defined, along with the problem itself.

[88] Sprevak (2009) identified this issue early on, noting that functionalism can lead to a particularly "radical" conception of extendedness in which we can be committed to saying these external things do compute representations, a conception which, as he noted, seemed too strong, which left a choice between rejecting functionalism or rejecting extendedness. With the nuance between causal and constitutive roles, this issue seems to be resolved.

Where do the "external information-bearing" structures come from? And how is it that these structures bear the relevant information? In paradigmatically cognitivist style, there is no accounting for this particular cognitive situation. If such things are taken for granted, then nothing needs to be defined, in which case the functionalist problem-solving sense of extendedness makes good sense.

So, from the perspective of this thesis, the existing conception of extendedness is distinctly cognitivist because it takes a problem-solving orientation on extendedness, an orientation that arises from taking the cognitive situation as already-given. We can now turn to how "thinking with post-cognitivism" might conceive and define extendedness where the question is, how do tools and extensions define our domain of awareness, action and interaction – how do they contribute to generating our "cognitive situation"?

Consider the effects of skiing equipment in winter mountain terrain. Here the clothing as well as the skis all count as extensions as they extend a human's basic capacities – mobility, thermoregulation, and survival, generally. A problem-solving account of these extensions describes how skis, waterproofs and insulating layers enable us to solve the "problem" of, say, getting down the mountain in a relatively optimal way – it's faster, warmer, and in many cases more fun than without these things. However, whilst this is all definitely true, it commits the standard move of assuming the situation and what the "problem" actually is (from the perspective of the agent in question). It thus ignores the significance of how the situation was generated in the first place. Humans do not just turn up in winter mountain terrain with these extensions. It is with these extensions that such choices become possible. These extensions open up and define whole new worlds of possible awareness, action and interaction. The extensions redefine the situation for the agent in the way that they generate possibilities. The situation of being protected from cold and wet, and having means of getting down the mountain is entirely different from one in which the agent is there without those things. We are not led to an awareness of the significance of our extensions in this way with a problem-solving orientation on them. All of this equally applies to extensions, whether we are talking about skis, submarines, rocket ships, printing presses, the internet, artificial intelligence, or something else.

Now, the way in which our extensions define new domains and worlds makes for an important segue to the next E, embeddedness. There was once an evolutionary conception of humans of us as predators at the top of a food chain, but another growing position in evolutionary theory claims that humans are "niche-constructors" (Laland et al. 2016). The notion of niche-construction is very much like the notion of "problem-defining". A niche can be physical and material, as well as social and memetic, in which case it concerns the ecosystem of ideas being selected for. With our technologies and extensions we are modifying our environment and are building an ecological niche at planetary scale. Further, it seems to be the ecological niche to which we are now adapting. As the capacities of our technologies scale, so too do our capacities to modify and define our niche. *There is thus a feedback loop between our extendedness and embeddedness.* This feedback loop is, again, not something we are led to notice with a problem-solving view of these notions.

An ethical dimension also comes out of a "problem-defining" view of extendedness when we notice that there is "no free lunch". That is, every extension takes something, costs something. As we give more and more over to our extensions, and now even speak for example of according Rights and Agency to our machines[89], we appear to be creating a niche for ourselves in which we are increasingly in service to the very tools we built to serve our ends. This ethical dimension to our tools is important to understand, regardless of what side we come down on. That is, whilst it is unclear in the last whether our extensions are generally, or in particular cases, really a morally good or bad thing, tools

[89] For example, Gunkel's (2018) *Robot Rights* book.

differentially affect our individual and collective agency – think of the global moral impact of the scientific and industrial revolutions in Western Europe – and we do not get that awareness from a "problem-solving" reading of extendedness. The "no free lunch" concern is ancient. Plato was famously concerned about the cost to our cognitive capacities of the shift from an oral to written tradition:

> "For this invention will produce forgetfulness in the minds of those who learn to use it, because they will not practice their memory. Their trust in writing, produced by *external characters which are no part of themselves,* will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding, and you offer your pupils the appearance of wisdom, not true wisdom." (Gleick 2012: 30) (italics mine)

Concerning more modern extensions, here is Daniel Schmachtenberger:

> A GPS device on your phone is designed to get you where you need to go, and it does that. It was not designed to weaken your sense of direction and make you dependent upon it to feel safe in urban or rural areas. Yet it also does that. What value is there now in having a good sense of direction or in being able to give and remember directions and locations? (Schmachtenberger 2022)

The significance of the "no free lunch" idea in the context of the extendedness of cognition is that the flip side of extending our cognition is "outsourcing" capacities to our extensions. Plato's concern was that such outsourcing might lead to the degeneration of our cognitive capacities.

In the next subsection on the final 'E' to be discussed – embeddedness – the importance of interpreting the 4E's of cognition in terms of how they define a condition of cognition will be further explored, with a continued connection to the consequences for how we might similarly understand the effects of our technologies.

5.3.3. Embeddedness

The notion of embeddedness does not seem to be discussed as independently as the others. Where it is discussed, it is often by way of comparison or distinction from extendedness (Shapiro 2019: 237-240). Recall that there are both causal and constitutive ways of understanding extendedness, but embeddedness is understood only in terms of the *causal* effects of "extrabodily" or environmental features in solving a problem (Kiverstein 2018). In this understanding, features of the environment causally contribute to the solution of a problem or completion of a task but are not processing representations so do not count as *constituting* "mental" activity. The claim of embeddedness therefore is that there is a strong *dependency* relationship between the agent and their environment, but that the "where" of the mental and cognitive activity is still something inside the brain (Kiverstein 2018: 1). The line defining what counts as cognitive is still the head, or at most the body. Beyond this, the distinction from extendedness is not very strong because both concern the involvement of "extrabodily" apparatus.

The idea is that "[t]he cognitive "load" that a task requires can be reduced when the agent embeds herself within an appropriately designed physical or social environment…" because some of the effort of cognition can be "offloaded" to the environment (Shapiro and Spaulding 2021).

Consider again the example of the bartender discussed in the previous section on extendedness. Understood from the perspective of *extendedness*, when the bartender lays out in order the specific glasses for each drink, they are "offloading" some of the cognitive load of remembering all the drinks to the "information-structure" of the environment. In this way the "tool" extends the minds functional capacity by representing some of the information necessary to process the drinks order – the "information-bearing load on his working memory is thereby significantly lightened." (Kiverstein 2018: 21) What representation in tools looks like is hard to say. On the one hand, with something like an abacus, it seems admissible to say that the counters represent numerical values, but with other things, like the bartender's glasses, it may seem more appropriate to say that the representations are in the bartender's head, and the glasses are mere placeholders of sorts. Again, some theorists of extendedness are of a position that this functional extension is constitutive of the cognitive process, whilst others maintain that the ordering of the glasses is merely causally involved in effective cognitive processes still located in the brain.

So, for extendedness, there is debate about whether it is one or the other, but, and this is perhaps the key difference, for *embeddedness*, the functional role of the environment is restricted to a causal role, a "noncognitive environmental prop" in the words of Wheeler (2010: 2).

> "In all of these cases, the cognitive capacities of an individual are enhanced when provided with the opportunity to interact with features of a suitably organized physical or social environment." (Shapiro and Spaulding 2021)

To the extent that the discussion of embeddedness distinguishes itself from extendedness primarily in terms of whether the functional role of the "suitably organized physical or social environment" is causal or constitutive, for the purposes of this thesis, there is not actually a great difference between them because both continue to speak in terms of tools or features of an environment contribute in some way to problem-solving.

For most theorists in this space, the difference is *not* small though, being no less than a matter of the metaphysics of mind (Sprevak 2010; Kiverstein 2018). However, in the view of this thesis, both sides are alternative positions on the same underlying question about how a mind *solves* a problem. Embeddedness becomes a context of yet another discussion of the problem-solving capacities of minds, and so is not as progressive as might initially be felt. What becomes excluded from the discussion then is the question of how the agent came to be embedded in the particular situation in the first place. Such an understanding of embeddedness takes the cognitive situation as already-given, leaving us with an impoverished understanding the significance of the environment for the agent from their perspective.

Therefore, alongside considering how features of our environment help us solve problems, a relevant question deserving of more attention is how the environments in which we are embedded contribute to *defining* our cognitive situation. Such a question is a point of departure for "thinking with post-cognitivism" about embeddedness, and becoming aware of the feedback loop between the environment and niche in which we are embedded, and the environment and niche which we create, as well as the effect of technological extensions in this loop.

These matters are not worked out in the literature. The discussion which follows here is therefore speculative, but offers insight which, I think, speak to the merit of a problem-defining conception of the 4E's and, by extension, its distinction from problem-solving cognitivism.

5.3.4. Ecological Embeddedness and "Superintelligence"

From a problem-defining perspective, being embedded in an environment means being bound by its constraints. "Problems" will therefore be problems related to the constraints of that environment. In this way, embeddedness identifies a logic of ecological constraint on a cognitive agent in the sense that what is possible for an embedded agent is partly ecologically defined.

This kind of conception of embeddedness in terms of the ecological constraints is language that is familiar to the field and context of "ecological psychology" (Lobo et al. 2018) where the "constraints" of environment are referred to as "affordances", a term pioneered by James J. Gibson in (1966, 1979) and still used today (e.g.: Chemero 2011, Rietveld and Kiverstein 2014). There is a close relationship between ecological psychology and the enactive, "4E" theory of cognition, with some attempts to integrate them into a single theoretical framework rooted in their shared "post-cognitivist" commitments (Rietveld et al. 2018, McGann et al. 2020, Heras-Escribano 2021). Here for example is McGann et al. (2020):

> "Despite these shared commitments and other apparent resonances between the approaches, communication between these two groups of researchers has been surprisingly sparse, and collaboration more rare still. Though several authors… have recommended some form of integration between them, just what such an integration would entail, and whether it might even be possible, has not been worked out in detail." (McGann et al. 2020: 1).

The notion of affordances in particular speaks to a "problem-defining" interpretation of embeddedness: "by "affordances" we mean the possibilities for action provided to us by the environment" (Rietveld et al. 2018: 2). The way in which the environment defines "possibilities for action" – a.k.a. affordances – invites an understanding of the environment of an embedded agent as defining the "problem-space" or situation for agent.

The understandings of embeddedness are not identical. The affordances of an environment speaks of the possibilities for action whilst "problem-space" has a normative or value-laden sense to it, one not present in mere possibility. One point of discussion in the thinking of ecological psychology which has its counterpart in the enactive theory, is the question of whether affordances are mind-independent features of an environment, or in what way there is some mutual coupling in an organism-environment system. Recall from the chapter on enactive theory of mind that the enactive conception of cognition of "sensemaking" is very much mind-dependent such that organisms perceive what is relevant for them to perceive as organisms trying to survive, and with perceptual capacities particular to their embodiment. Lobo et al. (2018) initially describe affordances in a way that expresses some similarities to enactive models of cognition:

> Gibson claimed that perceiving affordances is perceiving ecological meaning…which is perceiving how the surroundings are related to the agent's capacities. The idea of affordance shows that an organism does not perceive an objective, value-free, physical world in which meaning is imposed, as in Gestalt theory… We do not create affordances when we perceive them…they already exist in the system as constant relations between organism and environment. The detection of information amounts to affordance perception, so affordances are meaningful objects of perception in an organism-environment system. (Lobo et al. 2018: 7)

Having said this, they soon after note that there is not a single definition of affordances, only, really, "authors who claim that affordances are properties of the environment that are complemented by aspects of organisms, while there are other authors claiming that affordances are properties of the organism-environment system." (ibid)

In any case, whether the affordances are features of the environment, or are some sort of relational property, it may be said that being embedded in an environment defines the "possibilities for action" of an agent and in this way defines the possible kinds of problems an agent is capable of encountering. This is quite a different conception and interpretation of the role of the environment to previously mentioned ones in which it is understood as either causally or constitutively contributing to the execution and solution of a cognitive process.

If the ecological-constraint - problem-defining reading of embeddedness is taken seriously, it affords some important reflections on the notion of "superintelligence". As a reminder, here is how Bostrom defined it: "…we use the term "superintelligence" to refer to intellects that greatly outperform the best current human minds across many very general cognitive domains" (Bostrom 2014: 52).

Now, I have claimed in this thesis that superintelligence is a distinctly cognitivist concept. In the context here of ecological embeddedness, this is all the clearer. To put it provocatively, superintelligence makes about as much sense as an infinitely tall tree. The ecosystem and environment in which an intelligence is embedded place ecological constraints on the scaling of that intelligence. Even if the structural integrity of its organic matter did permit it to grow arbitrarily tall, (in the way that cognitivists imagine of the infinite scaling of intelligence), which it does not[90], the development of a tree is ecologically constrained by the energy and nutrient cycles in which it is embedded. A tree cannot extract arbitrary or exponential amounts of energy from its environment to fuel its growth. Believing that computational intelligence can be more or less infinitely scaled exposes a lack of awareness of the ecological embeddedness of agents, even artificial ones. A notion of (instrumental) intelligence rooted in a conception of cognition as computation allows in theory for infinite scaling – superintelligence – but a notion of intelligence rooted in cognition as autopoietic sensemaking, whatever such a notion of intelligence might amount to, is not possible. This distinction is worth further exploration beyond this thesis.

The claim I have repeated throughout this thesis is that the cognitivist model of cognition takes the cognitive situation as already-given. Reconsidered in light of this discussion of embeddedness, if the

[90] See Geoffrey West's *Scale: The Universal Laws of Growth, Innovation and Scale: Sustainability in Organisms, Economies, Cities and Companies* for a discussion of scaling laws in different kinds of complex adaptive systems. The basic point I am making here is that for any kind of cognition based in the biological processes of autopoiesis, infinite scaling of that cognition (sensemaking) is not possible because of the ecological constraints on the individual.

role of the environment is not included and accounted for in the model, then there is nothing constraining the possible scaling of intelligence, bar basic limits of physics.

The importance of including the constraints of ecological embeddedness in our models is relevant more broadly too for the ecological world in which we build AI systems. Kate Crawford (2021) identifies a variety of ways in which the embeddedness of our computational systems has been more or less totally ignored, from the extractive practices of lithium mining and the inhumane labour practices required to mine and produce parts for AI systems which are lauded for their mythic levels of "intelligence".

> "Advanced computation is rarely considered in terms of carbon footprints, fossil fuels, and pollution; metaphors like "the cloud" imply something floating and delicate within a natural, green industry. Servers are hidden in non-descript data centres, and their polluting qualities are far less visible than the billowing smokestacks of coal-fired power stations. The tech sector heavily publicizes its environmental policies, sustainability initiatives, and plans to address climate-related problems using AI as a problem-solving tool…As Tung-Hui Hu writes in *A Prehistory of the Cloud,* "The cloud is a resource-intensive, extractive technology that converts water and electricity into computational power, leaving a sizeable amount environmental damage that it then displaces from sight." …" (Crawford 2021: 41-42)

Crawford's work identifies the very concrete expressions of cognitivism in the AI and tech industry, bringing awareness to the way in which what is assumed in the model of mind, is actively excluded from sight in the development of AI tech. That is, the development of AI technology is itself embedded in an environment, one with its one logic and structural incentives, including more obvious economic structures, as well as less obvious geopolitical structures.

As Crawford details at length, the logics of computation are tied to control and pervade the body of AI (cognitivist) thinking and work. From the cybernetic heritage of control systems (Wiener 2013), to the militarised history of AI funding and development (Dyson 2012, Crawford 2021: chap. 6), to the colonial dynamics of control, enclosure and extraction in the slave labour, lithium mining, and ecologically and energy-intense data mining necessary to physically build AI systems (Crawford 2021), to a conception of intelligence as optimisation for well-defined – i.e. controlled – problems (Bostrom 2014), a basic cultural forensics reveals a logic of control infused at every junction, concrete and abstract, of AI development. Compare this to the chaos and complexity of living systems simply trying to stay alive in an ocean of agencies beyond their control. Enacting thinking with post-cognitivism: "what is the problem here really?"

Now, none of these points are revealed by an inquiry into embeddedness in terms of how an environment contributes to *solving* a problem. Moreover, if problem-solving is the frame, a cognitivist might even say that there is actually a basically infinite supply of atoms available in our environment that can be commanded into computation for "more or less any final goal". It takes "thinking with post-cognitivism" to make sense of even ideas like 4E cognition.

By way of transitioning to the following (and last) subsection on embeddedness and 4E cognition, note that whilst a problem-defining framing of embeddedness is important for making sense of features of our models of mind, it also reveals some unnerving things about the kind of ends towards which technological development is oriented, and the societal effects of such an orientation.

AI as a technology is, then, ecologically embedded in a landscape of structures. Socioeconomic and geopolitical features all combine to shape the structural incentives shaping how and what kind of technology is built[91].

This concludes the subsection on embeddedness, and with it, the discussion of the larger section on the 4E's of cognition and, in turn, this chapter on post-cognitivism. I will now summarise the ground covered.

5.4. Conclusion

In this chapter, the goal has been to characterise the post-cognitivist paradigm of cognition in terms of "problem-defining". I wanted to show how the enactive theory and 4E cognition, understood specifically from a problem-defining perspective, offer an account of the genesis of the cognitive situation. The significance of this accounting is that, in the post-cognitivist model, there is a story about why the agent is facing the particular cognitive demands and problems that it is, one that is not present in the cognitivist model.

To do this, I focused first on the enactivist theory of mind, supporting it afterwards with a problem-defining perspective on the other Es of 4E cognition. Some refer to the Enactivism as a paradigm itself, e.g. (Stewart et al. 2010), but not all who make sense of mind and cognition in "non" or "post"-cognitive ways subscribe to all its details, so I have been treating enactivism here as a theory, a subset of a larger paradigm. Nonetheless, much as the computationalist account of mind expresses the fundamental features I wished to highlight in the cognitivist paradigm, the enactivism similarly expresses much which is fundamentally shared across the new movement and, for my purposes, expresses a conception of mind in which "problem-defining" is also fundamental.

To lay this out, I identified four key concepts: life-mind continuity, autopoiesis, sensemaking, and dynamic co-emergence.

The life-mind continuity thesis is important for the way in which it situates mind research in a biological and ecological context. It asserts that the principles of life are continuous with the principles of mind such that, by coming to understand the principles of one, we may understand the other.

Having established the biological context, the notion of autopoiesis makes much more sense. I described it in terms of a Kantian "self-producing whole", one which exists "for and by means of itself". I complemented this with a discussion from complexity research centred on the notion of "constraint-

[91] An existing, concrete case study of this embeddedness is identified by the notion of the "attention economy", a phenomenon identified even before social media in (Davenport and Beck 2001). Digital platforms built by humans are embedded in structures that incentivise design of those platforms in ways that maximises financial return for the companies building those systems. The systems become designed to capture as much of the end user's attention as possible (enclosure), something achieved by means of chronic hyperstimulation. Once enclosed in the digital domain, our resource – the data we create by our activity – is extracted and sold so that said companies can invest in further research which will enable optimising attention and data extraction in a positive feedback loop for the financial interests of those building the systems, and a negative feedback loop for many end users, both psychologically as individuals, and collectively, in issues of social epistemology like "misinformation", and techno-political developments such as the involvement of Cambridge Analytical in the 2016 British "Brexit" referendum and the 2016 US Presidential Election, to name a few.

closure". Autopoietic systems are also complex systems, systems which produce the constraints necessary to generate the thermodynamic work to produce those very constraints, in an autopoietic "work cycle". From there I brought into light the significance of the distinction I was making between autopoietic and complex systems which define and produce themselves endogenously and those, like Kant's watch, Kauffman's cannonball, and Brooks' Roomba robot, which are defined exogenously, according to the perspective of an external observer. I suggested that this difference matters because, according to the enactive perspective, meaning is relative to an autopoietic organism trying to maintain itself – meaning concerns "my" world. I suggested that "meaningful perspective" is the product of endogenous dynamics, and that in cases of exogenous, "allopoietic" and "heteronomous" production, the meaning is always relative to an external observer. Autopoiesis is therefore significant for this thesis because it puts a capacity to define, and to define no less than the self, front and centre in an account of mind.

From autopoiesis, I then turned to speak about sensemaking, the mode of cognition particular to autopoietic beings. Sensemaking is the particular mode of interaction across the individuating threshold between an autopoietic being and its environment. It is adaptive discernment "for, and by means of, the self", a discernment and navigation of the features relevant to the maintenance of that autopoietic self. This is the frame of reference of sensemaking. Staying alive as an autopoietic individual is the context in which the significance of stimuli is interpreted. As such, there is a stake to getting it right, having a developed capacity to define what is relevant, what is going on, and what matters. I noted that for this reason, the sensemaking account of cognition is inconsistent with the Orthogonality Thesis. This capacity to define what is relevant is another step to legitimising a characterisation of post-cognitivism in terms of "problem defining".

Finally, I spoke about the dynamic co-emergence of self and world which is implied by the notions of autopoiesis and sensemaking. As an autopoietic being produces and actualises itself, it "brings forth" a world of features to which its unique sensory capacities are sensitive. Neither precedes the other. As a being emerges, necessarily the world it perceives emerges too. There is in this way a necessary fit between self and world. Not *the* world, but *its* world.

The concept of dynamic co-emergence does not obviously contribute to a characterisation of post-cognitivism but speaks to the conditions of possibility of autopoiesis and sensemaking, which do. Autopoiesis describes one "side" of the individuating threshold between self and world, sensemaking describes an action across the threshold *from the perspective of the organism* ("my" world), and dynamic co-emergence describes this process from a 3rd personal perspective – dynamics in "the" world.

Co-emergence also reveals some important and nontrivial implications of the account for our collective pursuits. Acknowledging that particular conditions of observation generate particular observations, how we achieve coherence and consensus across conditions of observation emerges as a serious question for postcognitivism. It touches on the fringes of a moral concern about what the post-cognitive paradigm, and the sense it makes of things, means more broadly. This connection is key to this thesis. It is a vital claim of this thesis that it matters what ideas we use to think about other ideas with and that cognitivist and post-cognitivist ideas take us in very different directions, certainly where AI is concerned. In the next and final chapter, I will explore the different paths they may take with respect to the question of whether AI can become more ethical than humans.

6. Can AI Become More Ethical Than Humans?

So, so you think you can tell
Heaven from hell?
Blue skies from pain?
Can you tell a green field
From a cold steel rail?
A smile from a veil?
Do you think you can tell?

- "Wish you were here", Pink Floyd

"It matters what matters we use to think other matters with; it matters what stories we tell to tell other stories with; it matters what knots knot knots, what thoughts think thoughts, what descriptions describe descriptions, what ties tie ties. It matters what stories make worlds, what worlds make stories."

- *Staying with the Trouble: Making Kin in the Chthulucene*, Donna Haraway

6.1. Introduction

In this chapter I will at last respond to the question of this thesis – can AI become more ethical than humans? With the context of the previous chapters, the claim of this thesis is that if ethics is a matter of "problem-solving", then AI may become more ethical than humans, but, to the extent that ethics involves "problem-defining", AI cannot become more ethical than humans. Given that existing work and thinking in AI is in fact cognitivist, there are therefore good reasons for thinking that existing work and thinking on AI faces nontrivial limits on its capacity to model and reproduce natural intelligence and cognition. This is of broader importance beyond the thesis, but, for the purposes of the thesis, it invites a question as to whether cognitivist models are really useful for comparing AI and humans. If cognitivist models face limits to their capacity to capture natural intelligence and cognition, it would be a category error to employ them in thinking, for example, about whether AI can become more ethical than humans.

Instead, the work of the previous chapters was to show that, as an "allopoietic" system, AI is not the kind of system which is capable of defining problems for itself in any other way than instrumental to a goal that has already been defined for it. This instrumentalised conception of cognition is both central to the success of AI systems and their capacity to significantly outperform humans at already-given problems.

The situation, then, is one in which there are two paradigmatically different ways of thinking about AI and whether it can become more ethical than humans, with AI research being an expression of just one of them. In the words of Donna Haraway's quote above, there are two different stories with which to tell a story here, two different kinds of thoughts with which to think thoughts, and stories with which

to make worlds, and, importantly, it matters how we proceed. In this chapter these two ways will be explored in a bit more detail, and I will conclude with a reflection on this choice.

Now, it is not otherwise obvious to intuition that AI could become ethical than humans, and the claim of this thesis is that there are at least two different responses available, respectively represented by cognitivism and post-cognitivism. The claim is that it depends on whether we think about the question with cognitivist or post-cognitivist ideas. If we think with cognitivism, it looks like AI *could* become more ethical than humans, and if we think with post-cognitivism, it looks like AI *cannot* become more ethical than humans, because AI cannot define problems, and this matters because existing thinking about AI is distinctly cognitivist.

In the following two sections, I present each case as an argument and discuss the premises and assumptions involved.

6.2. Conditional Conclusions

> "This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes more ethical than you?
>
> - (Bostrom and Yudkowsky 2014: 16-17)

In this section, I will formulate the conclusions of this thesis, and the premises that support them. At first glance, it is not obvious that AI could become more ethical than humans. What's more, our two paradigms of mind research, cognitivism and post-cognitivism, each lead to different conclusions about to whether AI can become more ethical than humans. Call this conclusion "C1":

> **C1: Our two paradigms of mind research, Cognitivism and Post-Cognitivism, lead to contradicting conclusions about whether AI can become more ethical than humans.**

In what follows, I present the premises in support of C1 by arguing for two conclusions, C2 and C3:

> **C2: If we take a cognitivist approach to cognition, we can be led to the conclusion that AI *can* become more ethical than humans.**
>
> **C3: If we take a post-cognitivist approach to cognition, we will conclude that AI *cannot* become more ethical than humans.**

These two conclusions jointly lead to C1. I will now take C2 and C3 in turn and identify the premises that support them. The total set of premises and conclusions will be shown at the end of this chapter for clarity. First, C2, concerning the cognitivist paradigm.

6.2.1. Thinking with Cognitivism: More Ethical in Terms of Problem-Solving

There are several important premises to highlight in order to get to C2, the claim that, if we take a cognitivist approach to cognition, then we can be led to the conclusion that AI *can* become more ethical than humans. In the course of Chapter 3 on AI and cognitivism, these premises were implicit and embedded in the presentation of the paradigm. I will now make them explicit. It begins with the cognitivist position on cognition as problem-solving.

P1: According to cognitivism, cognition is problem-solving.

I made the case for this by pointing to how the cognitivist model of mind assumes the basic cognitive situation of an agent with a problem to solve, in a particular environment. I claimed that cognitivism takes the cognitive situation as "already-given", pulling the term from (Varela et al. 1991:1). I argued that the significance of taking the situation as already-given is that it leads to a problem-solving conception of cognition because, once these features are already defined, all that remains for the cognitive process is to solve a problem: if the agent is already individuated, already supplied with a well-defined problem, and already embedded in an environment of features to navigate to a solution, problem-solving is all that is left for the model to describe.

I pre-empted this line of reasoning in chapter 2 with an exposition of the theoretical context of this thesis, namely, the existential risk of artificial superintelligence. I noted specifically that the notion of intelligence in use is one of "instrumental intelligence", which views intelligence as a generalised capacity to solve problems. I suggested that the implicit conception of mind involved in this context is rooted in a cognitivist conception of mind and moved with this view in chapter 3 to an in-depth consideration of the central concepts of cognitivism (computation, information, functionalism) to show how they conceptually prop-up this problem-solving account of intelligence and cognition. I showed how these concepts presume an already-given cognitive situation and thus support a problem-solving conception of cognition, and finished by making explicit the ways in which the operational conception of mind in the claims about the existential risk of superintelligence comes from the same cognitivist, problem-solving conception.

From the claim that the cognitivist view of cognition is problem-solving (P1), the next premise leading to C2 is that cognitivism takes ethics to be a problem, just like any other domain:

P2: According to cognitivism, ethical problems are problems just like any other.

This is a claim that I am taking as a plausible assumption. It seems plausible on the grounds that, if it doesn't matter for a capacity to solve a problem in what domain that problem exists, whether moral, political, mathematical, we can assume that the principles of intelligent problem-solving apply also to the domain of ethics. In this light, ethics is just another domain or "environment", distinguished by the features and problems particular to it, but otherwise a domain like any other. So, as long as the problem is already-given, there is nothing special about ethical problems as problems that means that an AI system could not in principle solve them. As long as the ethical problem and goal is well-defined, as cognitivist models of intelligence and cognition assume, then ethics is a matter of utility functions and

rational action-selection just like any other domain, with no necessary reflection on that goal, as claimed in the Orthogonality Thesis.

If there *is*, something special about ethics, it remains to be determined and accounted for, and that would complicate things for cognitivism. Cognitivism would have to explain both what was different about ethics, and then it would have to explain how agents in its problem-solving model engage this different kind of thing.

Strictly speaking, P2 is a claim about ethical problems, and not ethics itself, so to be as explicit as possible, it is important to add the following premise:

P3: Plausibly, according to cognitivism, ethics is a variety of problem-solving.

This more or less follows from P1 + P2 because, if cognitivism treats ethics as a domain of problems like any other, then cognition in that domain amounts, as per the cognitivist position on cognition, to problem-solving. I am treating P3 as an assumption about what the cognitivist position on cognition would say about ethics.

With these three in place, the next claim is that problem-solving is something AI can do:

P4: AI can do problem-solving.

This is also a premise that I am taking for granted. That AI is designed and built to solve problems is explicitly discussed in chapter 2 in which the contemporary "rational agent" paradigm of AI (Russell and Norvig 2010, Russell 2019) is centred on the principle of building machines that can rationally solve problems. (See section 2.1.2.1. in particular.) I also noted throughout chapter 3 that the historical development of the notions of computation, information, and functionalism was in fact rooted in the development of problem-solving machines.

What premises 3 and 4 together imply is that "AI can "do ethics" in the sense that, if ethics is a variety of problem-solving (P3), then that is something AI can do (P4). This becomes another premise:

P5: According to cognitivism, AI can do ethics.

Premise 5 concerns the scaling of this capacity.

P6. According to cognitivism, because AI exceeds human capacities for problem-solving in more and more domains, it may also exceed human-level cognition in the domain of ethics.

Via this series of premises we may be led to C2, the conclusion that, according to cognitivism, AI could in principle become more ethical than humans. Again, the sense of P5 and P6 here is that being ethical amounts to performance on well-defined ethical problems.

The idea is that we have only to look at AI systems' proven performance across domains at solving problems that are defined and provided such that we would say AI is in fact exceeding human-level cognition at solving those problems, and then extend that to the domain of ethics. Again, there is no guarantee that AI will in fact be able to solve any given ethical problem. Some problems, however well-defined, may not be solvable for such a system.

Generally though, as we continue to observe superhuman performance in AI systems across domains, we can infer that performance in the domain of ethical problems would also reach superhuman levels. However, one objection is that the same questions about getting from instrumental to general intelligence apply here (see (Müller and Cannon 2022)), namely, just because AI might be superhuman in the domain of ethics, it might still lack a more general intelligence. That is, much as the superhuman performance of AI is currently only domain-specific and not yet "general", the situation as I have explored it in this thesis may amount to something similar: the AI in question which achieves superhuman levels of ethical problem-solving may, for its restriction to the domain, be little more than an ethical calculator, superhuman at solving problems in the domain, but incompetent elsewhere. The moral status of such a system is as yet undefined.

In any case, whether superhuman problem-solving in the domain of ethics genuinely amounts to "being more ethical than humans" is worthy of further research. For now, the concern is that it is the conclusion of the cognitivist paradigm. I will now turn to the conclusion and premises of the post-cognitivist paradigm.

6.2.2. Thinking with Post-Cognitivism: More Ethical in Terms of Problem-Defining

The post-cognitivist conclusion is more or less the negation of the cognitivist conclusion:

> **C3: If we take a post-cognitivist approach to cognition, we will conclude that AI *cannot* become more ethical than humans.**

There are a few important premises to make explicit in order to arrive at the conclusion. Like with cognitivism, the first is claim about the post-cognitivist view of cognition:

> **P7: According to post-cognitivism, cognition involves *both* problem-solving and problem-defining.**

Recall that, whilst I characterised post-cognitivism here as a problem-defining account of cognition, I pointed out that problem-defining was the *distinguishing* feature of the paradigm, compared to cognitivism. Post-cognitivism does also include problem-solving. The problem-solving aspects of post-cognitivism are therefore features I took for granted in this thesis, focusing on the phenomenon of

problem-defining. Chapter 5 on post-cognitivism began with a section defining problem-defining", and then discussed both the enactive theory of cognition and the 4E theory of cognition to argue for how the theories support the problem-defining characterisation.

The next premise is a claim about what problem-defining means:

P8: Problem-defining is "an individual's endogenous orienting discernment of situation and meaning"

This is the definition that was offered in the beginning of chapter 4 on post-cognitivism. "Endogenous orienting discernment" is the root of "problem-defining" as understood in this thesis. How it works is best understood in terms of the enactive theory of mind, and the notion of autopoietic sensemaking in particular. Autopoiesis is necessary. This is premise 9.

P9: Problem-defining requires autopoiesis.

This is discussed in full detail in chapter 5. For summary here, recall that the process of autopoiesis is the continuous, open-ended self-production of an organism. The process of an organism autopoietically enacting itself means it is producing the conditions for its body to continuously produce itself. This body becomes a basis of orientation in the world, and the thermodynamic and metabolic precarity of the embodied organism become the condition as which the organism interprets the meaning of the features it perceives. Recall also the "dynamic co-emergence of self and world" in which, the enactive organism and the world it perceives co-emerge. The features the organism encounters are transcendentally tied to the enaction of the organism – it perceives only what it is capable of perceiving, and what it is capable of perceiving is a function of its autopoietic enaction.

The discernment of situation and meaning are therefore tied to the autopoietic process of an organism enacting itself. Problem-defining requires autopoiesis.

The next premises are claims about whether AI can fulfil this requirement.

P10: AI is not autopoietic.

This too is a technical point, but hopefully clear now. Autopoietic systems are living systems and AI, in current incarnations, is not a living system[92], therefore AI is not autopoietic.

P11 (from P9+P10): AI is not capable of problem-defining.

[92] It remains an open question whether this is a hard line or whether, in time, we might find a means of building AI with biological materials in a way that reproduces something like "artificial autopoiesis".

Combining P11 and P7 (According to post-cognitivism, cognition involves *both* problem-solving and problem-defining) we get the following:

P12: (from P6+P10): According to post-cognitivism, AI systems cannot have cognition.

P13: Ethics, in the sense relevant for the dissertation, requires cognition.

Premise 13 is something that has been implicit in the thesis. It is an assumption I am taking for granted.

Because in the post-cognitivist paradigm AI does not have cognition (P11), any connection with ethics must be made explicit in order to claim that AI cannot, as a consequence, become more ethical than humans, and this is what the following P14 premise does. It is a premise that I am treating as a plausible assumption.

P14 (from P12+P13): According to post-cognitivism, AI systems cannot be ethical and, therefore, cannot become more moral than humans.

From this final premise we are led to the conclusion, C3, that, according to post-cognitivism, AI cannot become more moral than humans.

C3: If we take a post-cognitivist approach to cognition, we will conclude that AI *cannot* become more ethical than humans.

By way of summarising, I will now put all this premises and conclusions together.

6.3. Summary and Discussion

I will now present all the premises and conclusions just discussed in order to summarise and discuss the significance of the findings. I begin with the main conclusion of the thesis, C1, and then list the conclusions and premises which lead to C1.

C1: The two paradigms of mind research, Cognitivism and Post-Cognitivism, lead to different conclusions about whether AI can become more ethical than humans.

C2: If we take a cognitivist approach to cognition, we can be led to the conclusion that AI *can* become more ethical than humans.

P1: According to cognitivism, cognition is problem-solving.

P2: According to cognitivism, ethical problems are problems just like any other.

P3: Plausibly, according to cognitivism, ethics is a variety of problem-solving.

P4: AI can do problem-solving.

P5. According to cognitivism, AI can do ethics.

P6. According to cognitivism, because AI exceeds human capacities for problem-solving in more and more domains, it may also exceed human-level cognition in the domain of ethics.

C3: If we take a post-cognitivist approach to cognition, we will conclude that AI *cannot* become more ethical than humans.

P7: According to post-cognitivism, cognition involves both problem-solving and problem-defining.

P8: Problem-defining is "an individual's orienting discernment of situation and meaning"

P9: Problem-defining ("an individual's orienting discernment of situation and meaning") requires autopoiesis.

P10: AI is not autopoietic.

P11: (from 9+10): AI is not capable of problem-defining ("an individual's discernment of situation and meaning).

P12: (from 7+11): According to post-cognitivism, AI systems cannot have cognition.

P13: Ethics, in the sense relevant for the dissertation, requires cognition.

P14: (from 12+13): According to post-cognitivism, AI systems cannot be ethical and, therefore, cannot become more moral than humans.

The conditional conclusion of this thesis therefore is that, if we think with cognitivism, it looks like AI can become more ethical than humans, and if we think with post-cognitivism, it does not look possible.

After all the work of the previous chapters, this may seem like a fairly underwhelming conclusion. However, existing views and work are AI is distinctly cognitivist and therefore make assumptions at a fundamental level about intelligence and cognition that are not universally shared within mind research. This is consequential in at least two ways.

Firstly, it demands questions about the legitimacy of AI systems as models of human, biological, and natural cognition. Secondly, it prompts metatheoretical questions about how seriously to take claims of superintelligence, orthogonality, and existential risk that make such assumptions about intelligence and cognition.

As we try to make sense of a technology which is blurring the line between tool and agent, we need sober discussion in both regards. In the final section before the concluding chapter of this thesis, I want to make some basic observations in this direction.

6.4. Choice, 2nd Order Cybernetics, and Undecidability

The work of this thesis is to establish two conditional claims about whether AI can become more ethical than humans. What this thesis does not do, alas, is offer a means for thinking about which paradigm of mind and cognition to go with, aware now of where each can lead us. The nature of the problem here is interesting and, given that something important could be at stake, I want to make some reflections before closing out this thesis. The basic issue seems to be that we need to make a choice, but do not have a basis for making that choice.

It is important to make a choice between paradigms here, I think, because each leads us down a different path in which we think about ourselves, our world, AI, and our relationship with it, in different ways. I have not done the work in this thesis to support any particular interpretation of what cognitivist or post-cognitivist visions of self, world, technology, and the relative place of things all look like. However, it should nonetheless be clear *that* there is a considerable difference of the various quality and degree of agency accorded to these things in each paradigm. In cognitivism, AI could be understood as something of a hyper-agent, or even hyper-moral agent, in which case it might be rational to give – abdicate – some kind of authority to AI. In post-cognitivism, it is just a technological extension, and the agency remains with the human[93]. This difference is morally significant and therefore demands some attention.

In the language of this thesis, these paradigms are "conditions of observation". Feminist epistemology and philosophy of science in particular have brought awareness to the ways in which our theories and paradigms influence our how we interpret observations and, moreover, our how we set up conditions of observation in the first place (e.g.: Barad 2007, Butler 2011, Haraway 2016). In this way, each paradigm generates particular observations when speculating about where AI might go. If we think about AI ethics with cognitivism, then we have to make sense of the moral significance of things like the Orthogonality Thesis and Superintelligence – for example, what is the moral status of an artificial superintelligence that could "solve climate change" but was indifferent to the goal, as per the orthogonality thesis. Or, if we think about the moral status of AI with post-cognitivism, conceptual exotica such as superintelligence and orthogonality do not come up. As far its moral status is concerned, from the perspective of post-cognitivism, AI is fundamentally different only by degree to any other machine because neither are living systems, meaning therefore any AI only has the moral status of a machine.

The directions available are quite different, but it is not clear on what basis to make a choice. The first reaction when faced with divergent or juxtaposed theories is to look to Truth to arbitrate between the two. That isn't possible here. Both paradigms are supported by considerable empirical science and

[93] Or "more-than-human"- whilst I do think that AI developments raise humanist concerns that seem too-easily parried from discussion of humans and machines in which what is important about humans is too-easily dismissed, I *do* want to avoid the human exceptionalism that blinds us the agencies of all the other creatures and forces of biology, ecology, and evolution. The expression "more-than-human" is an expression of posthumanism, a movement I discussed in an earlier (page 151). The notion of the more-than-human is the expression of this awareness that human agency is not the only agency on the planet, something for which Humanism has become subject to critique. Incidentally, posthumanism seems like it could be an ally of sorts to post-cognitivism for the way in which it (posthumanism) challenges the "already-given" or assumed frame of reference.

theoretical work. The choice of which to go with cannot, then, be made on the basis of which is true because both appear to be true.

Though I have spent the thesis differentiating them from one another, the paradigms are not as far as I can tell antithetical with respect to one other. A theoretical integration appears in principle possible. In the chapter on post-cognitivism, in section 5.2.3.5. in which I compared computation and sensemaking, I suggested that if autopoietic sensemaking is a circle, computation is an arc of that circle, describing a part of the larger process[94]. This may be a way of bringing them together, but what the question of whether AI can be more ethical then looks like from this metatheoretical perspective, remains to be discovered. And beyond this, we do not yet have an established metatheoretical structure for holding the theories together. So, it seems we cannot rely on even a more positive-sum metatheory to guide us here either.

One place I do think there is a source of guidance is in the movement of 2nd Order Cybernetics, and particularly some of the thinking of one of its main champions, Heinz von Foerster. It is not the place to dive into new material now, but the point is simple, and we have met the notion of cybernetics in a few places in the thesis[95], so not a huge amount needs to be said.

The important distinction is the following 1st order cybernetics was concerned with modelling systems. The properties of those systems are irrelevant for the moment. In the context of the sciences of mind, we can say that cybernetics was concerned with producing a model of the brain. 2nd order cybernetics, by distinction, was concerned with producing a model that *included* both the "observed system" (the brain) *and* the "observing system" (the agent modelling the brain) (Mead 1968, Dupuy 2000, von Foerster 2005). Basically, 2nd order cybernetics included the observer in their model. Thus, where cybernetics was concerned with goal-oriented systems, 2nd Order Cybernetics was concerned with humans building cybernetic systems – the "cybernetics of cybernetics" (ibid). The thinking was top-to-bottom infused with self-awareness about the goals and practice of the science, recognising that, creating models of ourselves is something we do to understand ourselves and therefore, a truly complete model of ourselves will have to include a model of that process. More simply, "it takes a brain to produce a theory of a brain" and so, an appropriate model of a human mind is not a model of an "observed system", but of an "observing system".

I don't want to get lost in the curiosities here. The point of interest for me is the interest in the notion of a self-aware science of mind for which 2nd order cybernetics called. Recall the notion of embeddedness from 4E cognition. Even without putting it through the move of 2nd order cybernetics to include the observer in the model, we recognise that our science of mind is embedded within cultural landscapes which influence how the science is conducted. When thinking about AI, and faced with questions about how to think about AI, it seems morally vital to consider the recursive effects of this thinking on culture and science. In the context of the discussion of the 4E's, I briefly mentioned a feedback loop between extensions and embeddedness, that, extended by our tools, we change the world in which we are embedded, and then are in turn changed by it. There seems to be something of a cultural and scientific path-dependence in which the path and paradigm we explore pulls us along with it, in a self-reinforcing dynamic. It is important to take responsibility for the effects of the ideas and theories we put out, because, again, as feminist philosophy of science has pointed out these ideas effect culture

[94] This line of inquiry is being further developed in a work-in-progress paper.
[95] See pages 77 and 111 in particular.

and science, and in order to take responsibility, it is crucial that our science of mind and AI develops a certain self-awareness.

Such a self-awareness in our sciences of mind would facilitate responding to the question of choice here between the paradigms. I suspect that the choice would still be difficult though because the question is, I think, in von Foerster's terminology "undecidable" (von Foerster 2003). For von Foerster, in a speech delivered on "Ethics and 2nd Order Cybernetics", he observes the distinction between "decidable" and "undecidable" questions, saying this on the matter:

> Only those questions that are in principle undecidable, we can decide. Why? Simply because the decidable questions are already decided by the choice of the framework in which they are asked, and by the choice of the rules used to connect what we label "the question" with what we take for an "answer." In some cases it may go fast, in others it may take a long, long time. But ultimately we arrive after along sequence of compelling logical steps at an irrefutable answer; a definite "yes," or a definite "no." But we are under no compulsion, not even under that of logic, when we decide on in principle undecidable questions. There is no external necessity that forces us to answer such questions one way or another. We are free! The compliment to necessity is not chance, it is choice! We can choose who we which to become when we have decided upon an in principle undecidable question. (von Foerster 2003: 5)

Arithmetic is a prototypical domain of decidable questions. The answers to questions of operations of addition and subtraction and so on are decided by the logic of the system in which they are posed, namely, arithmetic. "With which paradigm of mind research should we think about AI" is an undecidable question. It is not decidable with the logic of the system in which was expressed, whatever that is. Arguably, questions about whether human minds are machines and whether machines are minds are also undecidable. According to von Foerster's "metaphysical postulate", "only those questions that are in principle undecidable, we can decide" (ibid).

The vision is inspiring, and whether von Foerster is right about any of it is, funnily enough, undecidable. And so, the moral call-to-arms here is take responsibility for our sciences of mind and AI and their undecidable questions. We can begin by taking seriously the question of how to think about AI, and the possibility that it may be on us to make a choice.

7. Conclusion

> What a society chooses to measure is always tied into that society's shared narrative about the way the world works and what is valuable.
>
> - Zak Stein, *Education in a Time Between Worlds*

The findings of this thesis can be summarised as follows. To the question of whether AI can become more ethical than humans, what we think depends on the paradigm of mind with which we think. There are two major options, corresponding to the two major paradigms of mind research, paradigms which this thesis characterised as cognitivism and post-cognitivism. Because each has a simple, but fundamentally different conception of cognition, a different view of the moral potential of AI emerges. What emerges is this: according to cognitivism, AI can become more ethical than humans, whilst according to post-cognitivism it is not possible. Because for cognitivism cognition is "problem-solving", this is something AI can do, as long as the problem is defined for it. This means that, if "Ethics" as a task is somehow defined for AI, then AI can in principle surpass human "performance" at that task. There remains a caveat though about whether it would be an interesting phenomenon or just an "ethical calculator". In the post-cognitivist conception of cognition as sensemaking, cognition involves "problem-defining". From this view, becoming more moral than humans requires a capacity to define things, for which autopoiesis is necessary. This is a capacity AI does not have as an other-produced ("allopoietic") system. Hence, in this view, AI cannot become more ethical than humans.

The conditional conclusion of this thesis is the product of the distinction between problem-solving and problem-defining conception of cognition, characterisations which are idiosyncratic to this thesis and its purposes. They are not recognised terms in the fields to which they speak, but they afford a means of noticing two ways of thinking about cognition and AI.

This is important because the literature out of which the question of the thesis came – the literature on the existential risk of AI and superintelligence – is cognitivist in its assumptions about intelligence and cognition and displays no awareness or discussion of post-cognitivist research and conceptions of cognition. How in practice this should be done is unclear, but that it is of consequence is clear because we may be missing something very important about human and natural cognition in making cognitivist assumptions.

With that in mind, the aim of this thesis was to hold the paradigms together to both make sense of whether AI can become more ethical that humans, and to bring more attention and self-awareness to ways in which we do, and can, think about it. The juxtaposition of "problem-solving" and "problem-defining" characterisations is pragmatic to that end.

Whilst pragmatic and simple, the characterisations nonetheless seem to identify a deep distinction in our existing conceptions of cognition, with important implications for thinking about AI. After situating the thesis in the context of the existential risk of AI, the bulk of the work was in the two chapters on cognitivism and post-cognitivism, showing their fundamental conceptions of cognition to be of the respective "problem-solving" and "problem-defining" character. In both cases, the characterisation is due to the basic picture of the cognitive situation.

For cognitivism, the already-given cognitive situation involves an agent navigating an environment to solve a problem. Noticing that the model takes the cognitive situation already-given is key. The agent is already individuated, as is their problem, though not always by them themselves, and all of this happens in an environment that is also already-given. In this picture, cognition involves solving a problem whose significance is taken for granted, by an agent whose individuation is also taken for granted, in an environment whose features are too. With these variables already "defined", cognitivist cognition defines a process of solving a well-defined problem and, crucially, well-defined by an external source. Computation, the cognitivist conception of cognition, rooted in Turing's work on computable numbers and his eponymous machines, is the formal and formalisable term for this processing information to solve an already-defined function. Discerning the significance of the problem is not included in the cognitive process. It is "orthogonal" to the capacity to solve it, a capacity which can in principle be scaled beyond human-levels. In this way, cognitivism merits the characterisation of "problem-solving", and allows for the possibility that AI become more moral than humans, if the problem is adequately defined.

For post-cognitivism, explored in this thesis through the Enactive and 4E theories of cognition, the basic cognitive situation involves defining the very situation itself. The individuation of the agent, world, and problem are considered part of the problem of cognition. Rooted in biological autopoiesis, based on the Life-Mind Continuity Thesis, enactive cognition defines a process in which, as an organism autonomously produces itself, its capacities of perception develop and, in turn, what it is capable of perceiving also develop, meaning that what it perceives as "the world" also develops. Throughout this "dynamic co-emergence of self and world", the organism discerns and makes sense of what it is capable of perceiving in a way that is adaptive. Maintaining autopoietic integrity – surviving – becomes the frame of reference through which problems are discerned. In this way, "sensemaking" inherently involves an orientation and sensibility for what matters and what is relevant in a way that computation does not. "Problem-defining" happens at every stage. Autopoiesis involves "defining" the self, in the sense of individuating and maintaining the integrity of that self. And because the world for the organism emerges in this process, even defining its world can be said to be part of the cognitive process.

The difference then is that cognitivism assumes the player, the stage, and the "play" of cognition, (what I have been referring to as the basic cognitive situation), whilst individuating these things are part of the cognitive process for post-cognitivism. The logic of this thesis then was to notice and acknowledge the significance of this difference in our thinking about AI, so that we can be more aware of the way in which our speculations about the future of AI are conditional on our underlying conception of cognition and its assumptions. Again, existing thought and work in AI displays no awareness that its cognitivist conception of cognition and intelligence rests on assumptions that are not universally shared across mind research. AI research might therefore be missing something important. This matters both as far as we make speculations about the potential of AI, and as far as we leverage AI as a model of mind and method of mind research.

What this thesis does not do, is follow through the rest of the argument after the conditional. There is no follow-up stipulating which paradigm is "right", or which one we *ought* to be thinking with. Again, the main point of the thesis was to recognise that what we think about AI is conditional on what we think about AI as a mind, and that existing AI research makes certain assumptions in this regard. That said, in the last section, I offered some reflections on why it seems important to nonetheless take this question seriously.

AI displays apparent *features* of a cognitive system, both like ours and not. According to cognitivism it counts as a cognitive agent, but according to post-cognitivism it is not a mind or cognitive system because it is not living, autopoietic system. Their compatibility notwithstanding, as fundamentally

different ideas about minds, they afford equally distinct maps of the potential of AI. Cognitivism is the paradigm of mind that has emerged out of the effort to so precisely describe the various functions of humans that, in the words of the famed proposal for the Dartmouth Summer Research Project on AI, "a machine can be made to simulate it". It is therefore naturally well suited to describing the behaviour of a machine. The open question is whether this necessarily means it is also well suited to projecting the potential of AI. A "machine" is a model of mind, and existing AI is the most sophisticated expression of that model. The open question is, how good is the model? Humans both are and are-not like machines. Ultimately therefore, Superintelligence and existential risk are projections from a model which both does and does not fit, the same model which, according to this thesis, allows for the possibility that AI may become more ethical than humans.

When we notice the model of another paradigm, we can recognise the original model as such. The post-cognitivist model is not a model of a machine, but of a living system, the summary difference being that, where a living system produces itself (autopoiesis) "for, and by means of, itself", a machine is produced by, and for, something else. A computer does not build itself and navigate its world with a view to staying alive. It is built by a system of humans for a purpose given to it. In the model of living systems, accounting for this genesis and individuation of the agent are part of the model of cognition.

The aspiration of this thesis is to be found more in the fruit of holding these two paradigms together than its particular conditional conclusion statement. The parting impression is of a situation in which one model accurately describes machines at the expense of the living, and another which accurately describes the living, at the expense of the machines. Both are exclusionary.

The concluding statement of this thesis is, in the last, less significant than the general conclusion that any projection concerning the potential of AI is conditional. The "problem" then, if you will, is to be found, not in discerning whether AI can "become more ethical than humans", but in the question of how to make sense of the different ways of making sense of AI. Recognising that there are different ways to make sense of AI means we have to make sense of the impact of taking each path. The two paradigms tell us that both the similarities and the differences are profound. Moving forward from here, making sense of the different ways of making sense of AI will amount to acknowledging, understanding, and taking as much responsibility as possible for the effects of particular maps, models, stories and ideas about a technology that touches everything.

Bibliography

Adams, F. (2003). The Informational Turn in Philosophy. *Minds and Machines*, *13*(4), 471–501.

Adriaans, P. (2020). Information. In E. N. Zalta (Ed.), *The Stanford Encyclopaedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/information/

Aizawa, K. (2018). Critical Note: So, What Again is 4E Cognition? In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 116–126). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.6

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511978036

- (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, *28*, 15–26.

Arendt, H. (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*. Penguin UK.

Armstrong, S. (2013). General Purpose Intelligence: Arguing the Orthogonality Thesis. *Analysis and Metaphysics* 12: 68-84

Atkins, P. W. (1984). *The Second Law*. W. H. Freeman and Co.

Bak, P., & Paczuski, M. (1995). Complexity, contingency, and criticality. *Proceedings of the National Academy of Sciences*, *92*(15), 6689–6696. https://doi.org/10.1073/pnas.92.15.6689

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Bateson, G. (2015). Form, Substance and Difference. *A Review of General Semantics*, *72*(1), 90–104.

- (1979). *Mind and Nature: A Necessary Unity* (1st ed). Dutton.

Baudrillard, J. (2009). *The gulf war did not take place* (P. Patton, Trans.; Reprint). Power Publications.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B. S., Ó hÉigeartaigh, S., Martínez-Plumed, F., Price, H., Shevlin, H., Weller, A., Winfield, A., & Hernández-Orallo, J. (2018). Mapping Intelligence: Requirements and Possibilities. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017* (Vol. 44, pp. 117–135). Springer International Publishing. https://doi.org/10.1007/978-3-319-96448-5_13

Bitbol, M., & Luisi, P. L. (2005). Autopoiesis with or without cognition. *Journal of the Royal Society Interface*, *1*, 99–107.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Block, N. J., & Fodor, J. A. (1972). What Psychological States are Not. *The Philosophical Review*, *81*(2), 159–181. https://doi.org/10.2307/2183991

Brito, C. F., & Marques, V. X. (2016). Is There a Role for Computation in the Enactive Paradigm? In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 79–94). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_6

Brooks, R. A. (2002). *Flesh and Machines: How Robots Will Change Us*. Pantheon Books.

- (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M
- (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*(1–2), 3–15. https://doi.org/10.1016/S0921-8890(05)80025-9

Bostrom, N. (2016). *Existential Risk FAQ*. Existential Risk: Threats to Humanity's Future. Oxford University Future of Humanity Institute. https://www.existential-risk.org/faq.html

- (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- (2013). Existential Risk Prevention as Global Priority: Existential Risk Prevention as Global Priority. *Global Policy* 4 (1): 15–31. https://doi.org/10.1111/1758-5899.12002
- (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3
- (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, Vol. 9 (1).

Bostrom, N., & Cirkovic, M. M. (Eds.). (2011). *Global Catastrophic Risks*. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020

Bourgine, P., & Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life*, *20*, 327–345.

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 355–372. https://doi.org/10.1080/0952813X.2014.895108

Butler, J. (2011). *Bodies That Matter: On the Discursive Limits of Sex*. Routledge Classics.

Cannon, M. (2023). Humanise the Machine. *Phi Mag*, (Power).

- (2022). An Enactive Approach to Value Alignment in Artificial Intelligence: A Matter of Relevance. In V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence 2021. Springer Cham. https://link.springer.com/book/9783031091520

Carroll, L. (2009). *Alice's Adventures in Wonderland and Through the Looking-Glass*. Oxford University Press.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199207077.001.0001

Cappuccio, M., & Froese, T. (2014). Introduction. In M. Cappuccio & T. Froese (Eds.), *Enactive Cognition at the Edge of Sense-Making: Making Sense of Non-Sense* (pp. 1–33). Palgrave Macmillan UK. https://doi.org/10.1057/9781137363367_1

Carse, J. P. (2012). *Finite and Infinite Games: A Vision of Life as Play and Possibility*. Free Press.

Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science*, *12*, 325–359.

- (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*(3), 309–333.
- Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, *4*(4), 391–402.

Chemero, A. (2011). *Radical Embodied Cognitive Science*. MIT Press.

Chomsky, N. (1959). Review of B. F. Skinner's *Verbal Behavior*. *Language*, 35, 26–58. https://doi.org/10.4159/harvard.9780674594623.c6

Churchland, P. S., & Sejnowski, T. J. (1992). *The Computational Brain*. A Bradford Book.

Chrisley, R. L. (1994). Why everything doesn't realize every computation. *Minds and Machines*, *4*(4), 403–420. https://doi.org/10.1007/BF00974167

Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books.

Clark, A. (2008). Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind?*. *Philosophy and Phenomenological Research*, *76*(1), 37–59. https://doi.org/10.1111/j.1933-1592.2007.00114.x

- (2005). Intrinsic Content, Active Memory and the Extended Mind. *Analysis*, *65*(1), 1–11.
- (2001). *Mindware. An Introduction to the Philosophy of Cognitive Science*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Coeckelbergh, M. (2022). *Robot Ethics*. The MIT Press.

Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 155–179. https://doi.org/10.1080/09528130802319078

Davenport, T. H., & Beck, J. C. (2001). *The Attention Economy: Understanding the New Currency of Business*. Harvard Business Press.

Davidson, D. (1970). *Essays on Actions and Events: Philosophical Essays Volume 1*. Oxford, GB: Clarendon Press.

DeDeo, S. (2018). Information Theory. *Santa Fe Institute: Complexity Intelligence Report*, 15.

de la Bellacasa, M. P. (2017). *Matters of Care: Speculative Ethics in More Than Human Worlds*. University of Minnesota Press.

Dennett, D. (1991). *Consciousness Explained*. Back Bay Books: Boston, USA.

- (1987). *The Intentional Stance*. MIT Press.
- (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution* (pp. 129–151). Cambridge University Press.
- (2017). From Bacteria to Bach and Back: The Evolution of Minds. Penguin UK.

Di Paolo, E. (2018). The Enactive Conception of Life. In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 70–94). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.4

Di Paolo, E. A, Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor Life: An enactive proposal*. Oxford University Press.

Di Paolo, E. A., Cuffari, E. C., & Jaegher, H. D. (2018). *Linguistic Bodies: The Continuity between Life and Language*. MIT Press.

Di Paolo, E. A., & De Jaegher, H. (2022). Enactive Ethics: Difference Becoming Participation. *Topoi*, *41*(2), 241–256. https://doi.org/10.1007/s11245-021-09766-x

Di Paolo, E. A., Rohde, M., & De Jaegher, H. (2014). Horizons for the Enactive Mind: Values, Social Interaction, and Play. In *Enaction*. The MIT Press. https://doi.org/10.7551/mitpress/9780262014601.003.0003

Dodig-Crnkovic, G. (2022). Cognitive Architectures Based on Natural Info-Computation. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2021* (Vol. 63). Springer.

- (2020). Natural Morphological Computation as Foundation of Learning to Learn in Humans, Other Living Organisms, and Intelligent Machines. *Philosophies*, *5*(3), 17. https://doi.org/10.3390/philosophies5030017
- (2018). Cognition as Embodied Morphological Computation. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (2017) (pp. 19–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-96448-5_2
- (2016). Information, Computation, Cognition. Agency-Based Hierarchies of Levels. In V. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 139–159). Springer.
- (2014). Modelling Life as Cognitive Info-computation. In A. Beckmann, E. Csuhaj-Varjú, & K. Meer (Eds.), *Language, Life, Limits* (pp. 153–162). Springer International Publishing. https://doi.org/10.1007/978-3-319-08019-2_16
- (2011). Significance of Models of Computation, From Turing Model to Natural Computation. *Minds and Machines*, *21*(2), 301–322. https://doi.org/10.1007/s11023-011-9235-1
- (2008). Semantics of Information as Interactive Computation. In M. Möller, T. Roth-Berghofer, & W. Neuser (Eds.), *Proceedings of the Fifth International Workshop on Philosophy and Informatics*.

Dodig-Crnkovic, G. , & Burgin, M. (Eds.). (2011). *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation* (1st edition). World Scientific Publishing Company.

Dodig-Crnkovic, G., & Giovagnoli, R. (2013). *Computing Nature: Turing Centenary Perspective*. Springer Science & Business Media.

Dretske, F. I. (1983). Précis of Knowledge and the Flow of Information. *Behavioral and Brain Sciences*, *6*(1), 55–63. https://doi.org/10.1017/S0140525X00014631

- (1981). *Knowledge and the Flow of Information*. Cambridge University Press.

Dreyfus, H. L. (2007). Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology*, *20*(2), 247–268. https://doi.org/10.1080/09515080701239510

- Dreyfus, H. L. (2002). Intelligence without representation – Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, 1(4), 357–366. https://doi.org/10.1023/A:1021351606209
- (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.
- (1972). *What Computers Can't Do: The Limits of Artificial Intelligence* (Vol. 27, Issue 2, pp. 177–185). Harper & Row.

Dupuy, J.-P. (2000). *The Mechanization of the Mind: On the Origins of Cognitive Science* (M. B. DeBevoise, Trans.). Princeton University Press.

Dyson, G. (2012). *Turing's Cathedral: The Origins of the Digital Universe*. Penguin UK.

Edelman, S. (2008). On the nature of minds, or: Truth and consequences. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(3), 181–196. https://doi.org/10.1080/09528130802319086

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science*, 41(3), 253–259. https://doi.org/10.1016/j.shpsa.2010.07.009

Egbert, M. D., & Barandiaran, X. E. (2022). Using enactive robotics to think outside of the problem-solving box: How sensorimotor contingencies constrain the forms of emergent autonomous habits. *Frontiers in Neurorobotics*, *16*. https://www.frontiersin.org/articles/10.3389/fnbot.2022.847054

Feigl, H. (1958). The 'Mental' and the 'Physical'. *Minnesota Studies in the Philosophy of Science*, *2*, 370–497.

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics*, *7*(1), 1–11. https://doi.org/10.1007/s11571-012-9219-8

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

- (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2), 351–370.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548774.001.0001

- (1983). *The Modularity of Mind*. MIT Press.
- (1981). *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Bradford Books.
- (1975). *The Language of Thought*. Harvard University Press.
- (1968). *Psychological Explanation: An Introduction to The Philosophy Of Psychology* (Vol. 80, Issue 1, pp. 108–113). Random House.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fricker, M. (2007). Epistemic Injustice: Power and the Ethics of Knowing. Clarendon Press.

Froese, T., & Ziemke, T. (2009). Enactive Artificial Intelligence: Investigating the systemic organization of life and mind. *Journal of Artificial Intelligence*, *173*(3–4), 466–500.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford University Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception: Classic Edition* (Classic). Houghton, Mifflin and Company.

- (1966). *The Senses Considered as Perceptual Systems*. Bloomsbury Academic.

Giubilini, A., & Savulescu, J. (2018). The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence. *Philosophy & Technology*, *31*(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, *6*, 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0

Gleick, J. (2012). *The Information*. Fourth Estate.

Glock, H.-J., & Kalhat, J. (2016). Linguistic turn. In *Routledge Encyclopaedia of Philosophy* (1st ed.). Routledge. https://doi.org/10.4324/0123456789-DD3600-1

Gluckman, P. D., Hanson, M. A., & Low, F. M. (2019). Evolutionary and developmental mismatches are consequences of adaptive developmental plasticity in humans and have implications for later disease risk. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1770), 20180109. https://doi.org/10.1098/rstb.2018.0109

Gunkel, D. J. Press. (2018). *Robot Rights*. MIT Press.

- (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT

Hakli, R., & Seibt, J. (2017). *Sociality and Normativity for Robots. Studies in the Philosophy of Sociality*. Springer.

Halina, M. (2015). *Kinds of Intelligence*. https://static1.squarespace.com/static/614deb883727ee4f167d79e9/t/61829481405b1c5eb32e38fd/1635947651052/Halina+CUP.pdf

Haraway, D. J. (2016). *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press.

Hartley, R. V. L. (1928). Transmission of Information. *Bell System Technical Journal*, *7*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hayles, N. K. (1999). *How we became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. University of Chicago Press.

Heidegger, M. (1962). *Being and Time* (J. Macquarrie & E. Robinson, Trans.). Blackwell Publishing.

Heras-Escribano, M. (2021). Pragmatism, enactivism, and ecological psychology: Towards a unified approach to post-cognitivism. *Synthese*, *198*(1), 337–363. https://doi.org/10.1007/s11229-019-02111-1

Hernández-Orallo, J. (2017). *The Measure of All Minds*. Cambridge University Press.

Hidalgo, C. (2016). *Why Information Grows: The Evolution of Order, from Atoms to Economies*. Penguin Books.

Hoffman, D. D. (2019). *The Case Against Reality: How Evolution Hid the Truth from Our Eyes*. Penguin.

- Hoffman, D. (1998). *Visual Intelligence: How we create what we see.* W. W. Norton.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin & Review*, *22*(6), 1480–1506. https://doi.org/10.3758/s13423-015-0890-8

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An Eternal Golden Brain* (20th Anniversary). Basic Books.

Houkes, W. & Vermaas, P. (2010). *Technical Functions: On the Use and Design of Artefacts*, Dordrecht: Springer.

- (2004). "Actions Versus Functions: A Plea for an Alternative Metaphysics of Artifacts", *The Monist*, 87(1): 52–71. doi:10.5840/monist20048712

Hutto, D. D., & Myin, E. (2012). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

Jackson, F. (1986). What Mary Didn't Know. *The Journal of Philosophy*, *83*(5), 291–295. https://doi.org/10.2307/2026143

- (1982). Epiphenomenal Qualia. *The Philosophical Quarterly (1950-)*, *32*(127), 127–136. https://doi.org/10.2307/2960077

Jonas, H. (1966). *The Phenomenon of Life: Toward a Philosophical Biology*. Chicago: University of Chicago Press. Reprinted by Northwestern University Press, 200.

Kant, I. (2007). *Critique of Judgment* (N. Walker, Ed.; J. C. Meredith, Trans.). Oxford University Press.

Kauffman, S. A. (2019). *A World Beyond Physics: The Emergence and Evolution of Life*. Oxford University Press.

- (2000). *Investigations*. Oxford University Press.
- (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Kallestrup, J. (2006). The Causal Exclusion Argument. *Philosophical Studies*, *131*(2), 459–485. https://doi.org/10.1007/s11098-005-1439-x

Kersten, L., Deane, G., & Dewhurst, J. (2017). Resolving Two Tensions in 4E Cognition Using Wide Computationalism. In A. H. G. Gunzelmann (Ed.), *Proceedings of the 39th Annual Conference of Cognitive Science Society* (pp. 2395–2400).

Kirchhoff, M. D., & Froese, T. (2017). Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy*, *19*(4), 169. https://doi.org/10.3390/e19040169

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691133850/physicalism-or-something-near-enough

- (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Bradford Books.
- (1984). Concepts of Supervenience. *Philosophy and Phenomenological Research*, *45*,153–176. https://doi.org/10.2307/2107423

Kiverstein, J. (2018). Extended Cognition. In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 18–40). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.2

Kiverstein, J., Farina, M., & Clark, A. (2013). The Extended mind thesis. In D. Pritchard (Ed.), *Oxford bibliographies online. Philosophy* (pp. 1–20). Oxford University Press. https://doi.org/10.1093/OBO/9780195396577-0099

Klein, C. (2008). Dispositional Implementation Solves the Superfluous Structure Problem. *Synthese*, *165*(1), 141–153.

Krakauer, D. C. (2019). *Worlds Hidden in Plain Sight: The Evolving Idea of Complexity at the Santa Fe Institute, 1984-2019*. SFI Press.

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., & Ay, N. (2020). The information theory of individuality. *Theory in Biosciences*, *139*, 209–223. https://doi.org/10.1007/s12064-020-00313-7

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Penguin.

Legg, S., & Hutter, M. (2007a). A Collection of Definitions of Intelligence. http://arxiv.org/abs/0706.3639

- (2007b). *Universal Intelligence: A Definition of Machine Intelligence*. arXiv. https://doi.org/10.48550/arXiv.0712.3329

Legg, S., & Veness, J. (2011). An Approximation of the Universal Intelligence Measure (arXiv:1109.5951). arXiv. https://doi.org/10.48550/arXiv.1109.5951

Letelier, J.-C., Cárdenas, M. L., and Cornish-Bowden, A. (2011). From l'homme machine to metabolic closure: Steps towards understanding life. Journal of Theoretical Biology, 286(0):100 – 113.

Levin, J. (2021). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2021st ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/functionalism/

Lewis, D. (1980). Mad pain and Martian pain. In N. Block (Ed.), *Readings in the Philosophy of Psychology* (Vol. 1, pp. 216–222). Harvard University Press.

Lloyd, E., Wilson, D. S., & Sober, E. (2011). *Evolutionary Mismatch And What To Do About It: A Basic Tutorial*.

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The History and Philosophy of Ecological Psychology. *Frontiers in Psychology*, *9*.

Mancuso, S. (2018). *The Revolutionary Genius of Plants: A New Understanding of Plant Intelligence and Behavior*. Simon and Schuster.

Margulis, L. (1998). *Symbiotic Planet: A New Look at Evolution*. Basic Books.

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

Marr, D., & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry*. https://dspace.mit.edu/handle/1721.1/5782

Maturana, H. R. (1987). Everything is said by an observer. In W. I. Thompson (Ed.), *Gaia: A way of Knowing. Political Implications of the new biology* (pp. 65–82). Lindisfarne Press.

- (1970). Biology of Cognition. *Autopoiesis and Cognition: The Realization of the Living*, *43*(Boston Studies in the Philosophy of Science), 2–58.

Maturana, H. R., & Varela, F. J. (1980/2012). *Autopoiesis and Cognition: The Realization of the Living*. Springer Science & Business Media.

- (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding* (Revised). Shambhala.
- (1973). *De máquinas y seres vivos: Una teoría de la organizacíon biológica.* Santiago, Chile : Editorial Universitaria.
- (1970). Biology of Cognition. In H. R. Maturana and F. J. Varela, *Autopoiesis and Cognition: The Realization of the Living*, pp. 2-58. Boston Studies in the Philosophy of Science, vol. 43.

Maxwell, J. C. (1868). On governors. *Proceedings of the Royal Society of London*, *16*, 270–283.

Mayr, O. (1971). Maxwell and the Origins of Cybernetics. *Isis*, *62*(4), 425–444.

McCarthy, J., Minsky, M. L., Rochester, N., Corporation, I. B. M., & Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* 13.

McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence* (pp. 463--502). Edinburgh University Press.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, *1*(2), 185–196. https://doi.org/10.1007/BF00361036

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biology*, *52*(1/2), 99–115.

McGann, M., Di Paolo, E. A., Heras-Escribano, M., & Chemero, A. (2020). Editorial: Enaction and Ecological Psychology: Convergences and Complementarities. *Frontiers in Psychology*, *11*.

McLaughlin, B. P. (2006). Is role-functionalism committed to epiphenomenalism? *Journal of Consciousness Studies*, *13*(1–2), 39–66.

McLuhan, M. (2005). *Understanding Media: The extensions of man*. Routledge Classics.

Mead, M. (1968). Cybernetics of Cybernetics. In H. von Foerster, J. D. White, L. J. Peterson, & J. K. Russell (Eds.), *Purposive Systems*. Spartan Books.

Merleau-Ponty, M. (2002). *Phenomenology of Perception*. Routledge Classics.

Mitchell, M. Mitchell, M. (2019). We Shouldn't be Scared by 'Superintelligent A.I.' *The New York Times*. https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html

- (2009). *Complexity: A Guided Tour*. Oxford University Press.

Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, *372*, 179–191. https://doi.org/10.1016/j.jtbi.2015.02.029

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*, 18–21. https://doi.org/10.1109/MIS.2006.80

Mora, T., & Bialek, W. (2011). Are Biological Systems Poised at Criticality? *Journal of Statistical Physics*, *144*(2), 268–302. https://doi.org/10.1007/s10955-011-0229-4

Müller, V. C. (2016). "New Developments in the Philosophy of AI". In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_1

- (2007). Is There a Future for AI Without Representation? *Minds and Machines*, *17*(1), 101–115. https://doi.org/10.1007/s11023-007-9067-1

Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, *35*(1), 25–36. https://doi.org/10.1111/rati.12320

Müller, V. C., & Hoffmann, M. (2017). What Is Morphological Computation? On How the Body Contributes to Cognition and Control. *Artificial Life*, *23*(1), 1–24. https://doi.org/10.1162/ARTL_a_00219

Neander, Karen, 1991, "The Teleological Notion of 'Function'", *Australasian Journal of Philosophy*, 69(4): 454–468.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Newen, A., Bruin, L. D., & Gallagher, S. (Eds.). (2018). *The Oxford Handbook of 4E Cognition*. Oxford University Press.

Noë, A. (2005). *Action in Perception* (Vol. 102). MIT Press.

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *The Bell System Technical Journal*, *3*(2), 324–346. https://doi.org/10.1002/j.1538-7305.1924.tb01361.x

Ornes, S. (2017). Core Concept: How nonequilibrium thermodynamics speaks to the mystery of life. *Proceedings of the National Academy of Sciences*, *114*(3), 423–424. https://doi.org/10.1073/pnas.1620001114

Omohundro, S. M. (2012). "Autonomous Technology and the Greater Human Good". In Müller, V. (ed.) *The Risks of Artificial Intelligence.* Boca Raton: CRC Press Taylor and Francis Group.

- (2008). "The Basic AI Drives." In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence: Proceedings of the First AGI Conference.* 171: 483-492.
- (2007). "The Nature of Self-Improving Artificial Intelligence" Paper presented at Singularity Summit 2007, San Francisco, CA, September 8-9.

Parfit, D. (1984). *Reasons and Persons*. Clarendon Press.

Petersen, S. (2017). Superintelligence as superethical. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2. 0: New Challenges in Philosophy, Law, and Society* (pp. 322–337). Oxford University Press.

Piccinini, G. (2016). The Computational Theory of Cognition. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 203–222). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_13

- (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199658855.001.0001
- (2008). Computation without representation. *Philosophical Studies*, *137*(2), 205–241.
- (2007). Computing mechanisms. *Philosophy of Science*, *74*(4), 501–526.

Piccinini, G., & Maley, C. (2021). Computation in Physical Systems. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/computation-physicalsystems/

Piccinini, G., & Scarantino, A. (2010). Computation vs. information processing: Why their difference matters to cognitive science. *Studies in History and Philosophy of Science,* Part A, 41(3), 237–246. https://doi.org/10.1016/j.shpsa.2010.07.012 (2011). Information Processing, Computation, and Cognition. *Journal of Biological Physics*, *37*, 1–38. https://doi.org/10.1007/s10867-010-9195-3

Pickering, A. (2010). *The Cybernetic Brain: Sketches of Another Future*. University of Chicago Press.

Place, U. T. (1956). Is Consciousness a Brain Process? *British Journal of Psychology*, *47*(1), 44–50. https://doi.org/10.1111/j.2044-8295.1956.tb00560.x

Plato. (2015). *Theaetetus*. Aeterna Press.

- (2009). *Four Dialogues*. Wildside Press LLC.

Poggio, T. (2012). The Levels of Understanding Framework, revised. *Perception*, *41*(9), 1017–1023. https://doi.org/10.1068/p7299

Prigogine, I., & Nicolis, G. (1977). *Self-Organisation in Non-Equilibrium Systems*. New York: Wiley.

Putnam, H. (1988). *Representation and Reality*. MIT Press.

- (1975). The Meaning of Meaning. *Minnesota Studies in the Philosophy of Science*, 7, 131–193.
- (1967). Psychological Predicates. In *Art, Mind, and Religion*. University of Pittsburgh Press.
- (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions of Minds* (pp. 138–164). New York University Press.

Pylyshyn, Z. W. (1986). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

- (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132. https://doi.org/10.1017/S0140525X00002053

Ratcliffe, M. (2018). Perception, Exploration, and the Primacy of Touch. In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 280–300). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.14

Rietveld, E., & Kiverstein, J. (2013). A Rich Landscape of Affordances. *Ecological Psychology*, *26*(4), 325–352. https://doi.org/10.1080/10407413.2014.958035

Rosati, C. S. (2016). Moral Motivation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Roskies, A. L. (1999). The Binding Problem. *Neuron*, *24*(1), 7–9.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, Purpose and Teleology. *Philosophy of Science*, *10*(1), 18–24.

Rumelhart, D. E., & McClelland, J. L. (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.

- (1986b). On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models* (pp. 216–271). MIT Press.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach: International Edition* (3rd ed.). Pearson.

Savulescu, J., & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide* (pp. 79–95). Springer International Publishing. https://doi.org/10.1007/978-3-319-09668-1_6

Schmachtenberger, D. (2022). *Technology is Not Values Neutral: Ending the Reign of Nihilistic Design*. The Consilience Project. https://consilienceproject.org/technology-is-not-values-neutral/

Searle, John R. (1995). *The Construction of Social Reality*, New York: The Free Press.

- (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

Seibt, J. (2020). Robots and Artificial Intelligence: Posthumanism as Robophilosophy. In M. R. Thomsen & J. Wamberg (Eds.), *The Bloomsbury Handbook of Posthumanism* (pp. 289–304). Bloomsbury Academic. http://dx.doi.org/10.5040/9781350090507.ch-023

Seibt, J., Hakli, R., & Norskov, M. (2014). *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy*. IOS Press.

Shagrir, O. (2005). The rise and fall of computational functionalism. In Y. Ben-Menahem (Ed.), *Hilary Putnam (Contemporary Philosophy in Focus)*. Cambridge University Press.

Shanahan, M. (2016a). Conscious Exotica. Aeon. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there

- (2016b). The Frame Problem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2016th ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/frame-problem/
- (2010). Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds. Oxford University Press.
- (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press.

Shanahan, M., & Baars, B. (2005). Applying global workspace theory to the frame problem. Cognition, 98(2), 157–176. https://doi.org/10.1016/j.cognition.2004.11.007

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 1–55.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press.

Shapiro, L. (Ed.). (2014). *The Routledge Handbook of Embodied Cognition*. Routledge.

Shapiro, L., & Spaulding, S. (2021). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/

Sheldrake, M. (2020). *Entangled Life: How Fungi Make Our Worlds, Change Our Minds & Shape Our Futures*. Random House Publishing Group.

Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts. https://doi.org/10.1037/11256-000

Sloman, A. (2011). What's information, for an organism or intelligent machine? How can a machine or organism mean? In G. Dodig-Crnkovic & M. Burgin, *Information and Computation* (pp. 393–438). World Scientific Publishing Company. https://doi.org/10.1142/9789814295482_0015

- (1984). *The Structure of the Space of Possible Minds*. https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-space-of-minds-84.pdf

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, *68*(April), 141–156.

Soares, N. (2016). The Value Learning Problem. *Machine Intelligence Research Institute*, 7

Sorensen, R. (2003). *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind*. Oxford University Press.

Sperber, D. (2002). In defense of massive modularity. In E. Dupoux (Ed.), *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler* (pp. 47–57). MIT Press.

Sprevak, M. (2009). Extended Cognition and Functionalism. *Journal of Philosophy*, *106*(9), 503–527. https://doi.org/10.5840/jphil2009106937

Stewart, J. (2010). Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of life to Consciousness and Writing. In J. Stewart, O. Gapenne, & E. A. Di Paolo (Eds.), *Enaction: Toward a New Paradigm for Cognitive Science* (pp. 1–32). MIT Press.

Stewart, J., Gapenne, O., & Di Paolo, E. A. (Eds.). (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.

Taddeo, M., & Floridi, L. (2005). Solving the Symbol Grounding Problem: A Critical Review of Fifteen Years of Research. *Journal of Experimental and Theoretical Artificial Intelligence*, *17*. https://doi.org/10.1080/09528130500284053

Thompson, E. (2015). *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. Columbia University Press.

- (2007). *Mind in Life*. Harvard University Press

Thomson, W. (1874). Kinetic Theory of the Dissipation of Energy. *Nature*, *9*(232), 441–444. https://doi.org/10.1038/009441c0

Turing, A. (1950). Computing Machinery and Intelligence. *Mind,* LIX(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

- (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Tweedy, L., Thomason, P. A., Paschke, P. I., Martin, K., Machesky, L. M., Zagnoni, M., & Insall, R. H. (2020). Seeing around corners: Cells solve mazes and respond at a distance using attractant breakdown. *Science (New York, N.Y.)*, *369*(6507). https://doi.org/10.1126/science.aay9792

Uebel, T. (2021). Vienna Circle. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/vienna-circle/

Varela, F. (1991). Organism: A meshwork of selfless selves. In A. I. Tauber (Ed.), *Organism and the Origin of Self* (pp. 79–107). Kluwer Academic Publishers.

- (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, *34*: 72–87.
- (1979). Principles of biological autonomy. North Holland New York.

Varela, F., Maturana, H., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. Biosystems, 5(4):187 – 196.

Varela, F., Thompson, E. B., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Villalobos, M., & Dewhurst, J. (2018). *Enactive autonomy in computational systems*. https://philarchive.org/rec/VILEAI-3

- (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior*, *25*(3), 117–128. https://doi.org/10.1177/1059712317710496

von Foerster, H. (2005). Understanding Understanding: Essays on Cybernetics and Cognition. *Complicity: An International Journal of Complexity and Education*, *2*(1). https://doi.org/10.29173/cmplct8737

- (2003). Ethics and Second-Order Cybernetics. In H. von Foerster, *Understanding Understanding* (pp. 287–304). Springer New York. https://doi.org/10.1007/0-387-21722-3_14

von der Malsburg, C. (1981). *The Correlation Theory of Brain Function* [Departmental Technical Report]. MPI. https://web-archive.southampton.ac.uk/cogprints.org/1380/

von Neumann, J. (1958). *The Computer and the Brain.* Yale University Press.

von Uexküll, J. (1957). A stroll through the worlds of animals and men. In K. S. Lashley (Ed.), *Instinctive Behavior: The Development of a Modern Concept* (pp. 5–80). International Universities Press.

Vörös, S. (2023). On What is Always Before Our Eyes: The Uncharted Depths of Francisco Varela's Thought. *Journal of Consciousness Studies*, *30*(11–12).

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right From Wrong*. Oxford University Press.

Wallach, W., & Asaro, P. (Eds.). (2017). *Machine Ethics and Robot Ethics*. Routledge. https://doi.org/10.4324/9781003074991

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The Varieties of Enactivism. *Topoi*, *36*(3), 365–375.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*, 97–125.

West, G. (2017). *The Universal Laws of Growth, Innovation and Scale: Sustainability in Organisms, Economies, Cities and Companies*. Weidenfeld & Nicholson.

Wheeler, M. (2010). In Defense of Extended Functionalism. In R. Menary (Ed.), *The Extended Mind* (pp. 244–270). The MIT Press. https://doi.org/10.7551/mitpress/9780262014038.003.0011

- Wheeler, M. (2008). Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem. *International Journal of Philosophical Studies*, 16(3), 323–349. https://doi.org/10.1080/09672550802113235
- (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press.

Whitehead, A. N., & Russell, B. (1927). *Principia Mathematica*. Cambridge University Press.

Weijer, C. J. (2020). Chemotaxis: Active Degradation of Attractant Enables Optimal Maze Navigation. *Current Biology*, *30*(23), R1436–R1438. https://doi.org/10.1016/j.cub.2020.09.084

West, G. (2017). *Scale: The Universal Laws of Growth, Innovation and Scale: Sustainability in Organisms, Economies, Cities and Companies*. Weidenfeld & Nicholson.

Wiener, N. (2013). *Cybernetics or, Control and Communication in the Animal and the Machine* (2nd ed.). Martino Publishing.

Wilson, R. A. (1994). Wide Computationalism. *Mind, New Series*, *103*(411), 351–372.

Wittgenstein, L. (2001) *Philosophical Investigations*, Oxford: Blackwell.

Yampolskiy, R. V. (2016). *Artificial Superintelligence: A Futuristic Approach*. CRC Press.

- (2013). "What to Do with the Singularity Paradox?" In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 397–413). Springer. https://doi.org/10.1007/978-3-642-31674-6_30

Yampolskiy, R., & Fox, J. (2012). Safety Engineering for Artificial General Intelligence. *Topoi*, *32*(2), 217–226.

Yudkowsky, E. (2013). "Intelligence Explosion Microeconomics". (Technical report) Berkeley, CA: Machine Intelligence Research Institute.

- (2008). Artificial Intelligence as a positive and negative factor in global risk. In E. Yudkowsky, *Global Catastrophic Risks*. Oxford University Press. https://doi.org/10.1093/oso/9780198570509.003.0021

.