Replicating cognitive self-sufficiency to realize computational self-sufficiency in Artificial General Intelligence models.

Ivan Serwano Kyambadde

PhD

University of York

Philosophy

November 2023

ABSTRACT

There is compelling evidence that qualitative states are integral to realizing the type of autonomy exhibited in cognitive general intelligence. Cognitive agents leverage experiential qualitative states in forming knowledge of the world, adopting attitudes to the world, and processing responses to general problems in the world. This capacity for self-directed, adaptive problem-solving, that allows cognitive agents to solve both familiar and novel problems, serves as the main inspiration for Artificial General Intelligence (AGI). However, current AGI development focuses on computational models that largely ignore the role of qualitative states. If the goal of AGI is to realize artefacts with general intelligence that is functionally like cognitive general intelligence, then it should be critical to replicate the self-directed autonomy that qualitative states enable in cognitive general intelligence.

This research aims to resolve the question *'How can the roles of qualitative states in supporting the self-directedness of cognitive general intelligence be replicated in models of artificial general intelligence?'* The thesis identifies that qualitative states are pivotal in (1) facilitating the self-grounding of mental computation with underived meaning from the world. (2) Enabling self-creditable control of the computation that yields actions. These two roles are critical for accounting for the intentional actions that are the hallmark of the autonomy exercised by cognitive intelligence. The thesis delineates natural accounts of these two functional roles. From these roles, it extracts two replicable principles: the Analog Principle and the Internal Model Principle. These principles can be used to ameliorate computational models of intelligence with autonomy like cognitive general intelligence. To demonstrate this, the thesis describes a model of intelligence supported by these principles and assesses the model for the intentional agency that is definitive of cognitive general intelligence.

AUTHOR’S DECLARATION.

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Table of Contents

Chapter 1: Introduction.

> *"Then she read the following words, engraved upon the copper plates of the man's body: Smith & Tinker's Patent Double-Action, Extra-Responsive, Thought-Creating, Perfect-Talking Mechanical Man Fitted with our Special Clockwork Attachment Thinks, Speaks, Acts, and Does Everything but Live."*
>
> Tik-Tok of Oz (Baum, 1914, p. 73).

1.0. Popular tropes for intelligent artefacts.

The first technology called Tik-Tok that captured the popular fancy was a robot in Frank. L. Baum's popular Oz books. According to the popular tales, Tik-Tok was a one of its kind, copper automata. It had a rounded torso and the appendages, i.e., legs, arms, and heads, of a human. It was capable of thoughts, speech, and actions. Each one of these capabilities were powered by a separate wind-up mechanism. Once wound down, the capacities were shut down until they were rewound. Unfortunately, Tik-Tok was incapable of managing its own winding-up. It was dependent on its human minders to keep it recharged. Tik-Tok also differed from humans in another significant way. The books frequently reminded the reader that Tik-Tok was not alive and as a result it was incapable of feelings. This did not stop it from being truthful and loyal in the service of its owner, a service it executed well above human capacity on account of its super-human strength.

Tik-Tok was one of the earliest representations of an intelligent human-like automaton in popular literature. It took its first bow in the book Ozma of Oz from 1907 (Baum, 1907). Yet one can already observe in Tik-Tok several characterisations of human-like intelligence in artefacts that have remained persistent and pervasive even as the idea of human-like intelligence in artefacts has transitioned from a literary trope to real life research and development. Two of those enduring characterisations are central in motivating this research.

The first characterisation is the notion that intelligent artefacts will eventually outstrip human capabilities in certain or all tasks. This notion motivates by justifying the research. It is a characterisation that this research endorses. It highlights the importance of taking an interest in engineering for human-like intelligence in machines. If done responsibly and carefully, intelligent artefacts can be leveraged to improve human lives by solving problems that are beyond human capacity to solve.

The second characterisation is the notion that feelings, are superfluous to general intelligence. The author, Frank Baum goes to great length to stress that Tik-Tok has no feelings. There are three notions one can detect when the author states that the automaton did not possess feelings. The most obvious claim is that Tik-Tok was incapable of experiencing emotional states like joy or sadness. The second notion is that Tik-Tok did not have the capacity to experience sensations that mark out exteroceptive and interoceptive experiences such as touch (exteroceptive) or hunger (interoceptive). The third notion is that Tik-Tok did not have the capacity to attach hedonic preferences to either the emotional or sensory experiences. All these notions of feelings, the emotional and mood feelings like joy, sadness etc, the bodily sensational feelings like touch, and hunger, and the hedonistic experiences of pleasure or displeasure are experienced in humans as qualitative states.

It might be sensible to agree with Baum that qualitative states in the phenomenal form that they are experienced in humans might be dependent on the biochemical substrate of the mind, and that therefore it would be unlikely that Tik-Tok would have phenomenal experiences.

However, this should not preclude Tik-Tok from having states in other forms that mark out moods, perceptual, and affective experiences. To state that Tik-Tok had no feelings either is to state that feelings are superfluous to human-like intelligence. This idea that automatons with human-like intelligence will lack the capacity for feelings in any form has become a widely accepted trope. This notion presents a challenge. It is a characterisation that this thesis finds problematic and identifies as a potential obstacle to the success of engineering for human-like intelligence. The stance taken in opposition to this characterisation informs the central claim in this thesis and is the grounds upon which this research builds.

This introductory chapter will highlight both the motivation and the research agenda in more detail. It will then give a succinct description of the tone and approach adopted in this thesis, furnished with an explanation for these choices. Lastly it will describe the lay of the land.

1.1. Motivation.

An artefact with human-like intelligence will eventually leverage mechanical advantages to outstrip human performance in all or certain tasks. This is an obvious enough claim that it has become a standard feature in portrayals of intelligent artefacts. For the most part, this is a good thing. It is not hard to imagine how beneficial an artefact that exceeds human capacity can be if developed responsibly to address intractable problems that stand in the way of human flourishing. It is also a timely need. At the time of writing, the world seems beset with seemingly intractable problems, e.g., global warming, pandemics , social inequities etc. These problems appear to be solvable, but their solution appears to lie just beyond current human capacity. The promise of machines with artificial general intelligence is that they will work alongside humans to augment human endeavours. The responsible development of intelligent artefacts is potentially beneficial to humankind.

Intelligence that outstrips human performance is not only a key trope in our imagination of intelligent artefacts, but it has also been borne out by practical experience. Tik-Tok's labours were augmented by its mechanical, super-human strengths. It enabled it to leverage that strength to outperform its human contemporaries at certain tasks. This trope is supported by the success that has been achieved in artificial intelligence (AI) research to date. AI research and development has taken tremendous steps towards developing artificial narrow intelligence (ANI). Artificial narrow intelligence are artefacts or solutions powered by intelligent algorithms that are deployed to solve domain specific tasks. They are domain restricted in the sense that they are designed to effect one solution. Any learning acquired in the performance of this singular problem-solving function cannot be transferred to a different domain. Hence the word 'narrow' in artificial narrow intelligence. The word 'narrow' references that the artefact is only capable in a narrow domain. Yet, despite these shortcomings, ANIs can leverage certain advantages like machine resilience, dedicated computing resources, and alternative paradigmatic approaches to problem solving to outperform humans at certain tasks. Because of these advantages, ANIs are increasingly being deployed in society today where they have shown themselves capable of augmenting and improving human labours in significant ways.

The success of research, development and deployment of ANIs has spurred interest in artificial general intelligence research (AGI). Artificial general intelligence is a poorly defined concept. However, there is emergent agreement that it refers to a type of artificial intelligence that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks, much like human intelligence. Unlike narrow AI, which is designed to perform domain specific tasks, AGI would have the capacity to perform intellectual tasks across several domains like a human. Potentially, it will generalize knowledge from one area to another, reason, solve problems creatively, and adapt to new situations without needing specific training for each new task.

In short AGI research aims to realize machines like Tik-Tok that are general purpose machines with intelligence that is functionally comparable to human (or animal intelligence). In the same way that ANIs outperform humans in narrow tasks, AGI artefacts will eventually be expected to outperform humans in general problem-solving efficacy.

Given how transformative and beneficial ANIs have been on society, AGIs with their broader domain efficacy and flexible reasoning are expected to be exponentially more beneficial when, and if, they are eventually realised.

The potential value of a realized AGI artefact has attracted immense interest to artificial general intelligence research. A global survey found forty-five different R&D projects in thirty countries attempting to solve for AGI by 2017 (Fitzgerald, et al., 2020). These figures had grown to 72 AGI projects in thirty-seven countries by 2020 (Fitzgerald, et al., 2020). Experts in the field project that we are anywhere from ten years to a century away from realizing AGI artefacts (Ford, 2018), (Grace, et al., 2018). The varied prognostics suggest that there is still some work to be done.

The position championed in this thesis is that bridging the gap is contingent on resolving philosophy of mind problems at the foundations of AI theory. This research aims to contribute to artificial general intelligence research by addressing a potentially critical conceptual gap in AI theory. Mainstream AI theory appears to have adopted a position on general intelligence that excludes the role of experiential states in modelling general intelligence. This thesis aims to raise questions about this omission and identify functional roles of experiential states that should be accounted for in computational models of general intelligence. The thesis will define these functional roles in the form of principles that can be replicated to inform the AGI agenda.

1.2. Central Claim.

The central claim in this research is a response to the second characterisation that is frequently attached to the idea of artefacts with human-like intelligence. It is the notion that qualitative states that mark out internal, perceptual, and affective experiences are superfluous to modelling general intelligence.

Tik-Tok's first appearance was in 1907. Already, in this early representation of a generally intelligent artefact, the author Frank. L. Baum is tireless in reminding the reader that Tik-Tok does not have the capacity for qualitative experiencing. However, Tik-Tok's lack of qualitative experiences does not get in the way of its capacity for human-like intelligence.

The idea that feelings are superfluous to intelligence predates the realisation of any ANI artefact but has been reinforced by the success achieved in building ANIs. Since the time Baum introduced us to Tik-Tok, the AI agenda has progressed from literary concept to research and development and yielded very advanced ANIs. These ANIs are often referred to as weak AI. Weak AI is not designed to have the properties of a mind. That is to say that they are not designed to have experiential states that drive intelligent decision making. They are designed to operate within defined rules and parameters. As a result, weak AI is unlike human cognition and lacks its capacity for experiential states. Even though ANIs are designed without this capacity, they are very proficient at domain restricted tasks.

The success in engineering for ANIs offers one possible explanation of why feelings have been disentangled from the conception of intelligence in artefacts. Despite the fact that ANI's are developed with disregard for experiential states, they have been highly effective at narrow tasks. It is tempting to conclude that the same path that led to the success in yielding artefacts with narrow intelligence will yield to artefacts with general intelligence. However, this assumption is just as likely to be a first-step fallacy. The term 'first-step fallacy' was first introduced in relation to AI research by philosopher Yehoshua Bar-Hillel and mainstreamed by Hubert L. Dreyfus. It refers to the mistaken belief that limited early success is a guarantor for ultimate success (Dreyfus, 2012). In this case, success achieved in developing ANIs without regard for engineering for the role of experiential states s is not necessarily a guarantor that the same approach will yield artificial general intelligence.

A second factor that might account for the widespread belief that experiential states are superfluous to models of general intelligence is the timing of the initial breakthroughs in AI research. The first wave of AI researchers consisted of luminaries such as Alan Turing, John McCarthy, Marvin Minsky, Allen Newell, Herbert Simon, Claude Shannon, Edward Feigenbaum and Julian Feldman. They presented ideas that pushed AI research from the popular imagination into an emerging research field in the 1950s and 1960s. At that time, the predominant theory of mind was behaviourism. Behaviourism sought to explain intelligence in terms of its output i.e., behaviour, without regard to the internal processes and sates that accounted for behaviour. In keeping with this outlook, early AI research proceeded as if the task at hand was to model humanlike input-output patterns with little or no regard for the specific internal processes or states that accounted for humanlike output. Experiential states, their functional role, and the underlying states that realized them were considered inessential in explaining and modelling intelligence. The success of the AI agenda was considered dependent on optimizing systems to produce humanlike output from the right input.

The idea that experiential states and general intelligence can be disentangled appears to be at variance with the single indisputable example of general intelligence that is available to us. Cognitive general intelligence is the only consensus example we have of general intelligence.

Cognitive intelligence encompasses the mental abilities needed for acquiring knowledge, learning, reasoning, and solving problems. It allows an agent to make informed decisions to solve problems independently. This form of intelligence includes both the capacity to adapt to new situations and the ability to apply previously learned knowledge and skills to familiar tasks. All of this is accomplished with an autonomy that is essential, because it locates the credit for intelligent actions in the cognitive agent itself.

Even under a cursory examination, it is hard to ignore that experiential states are writ large in cognitive general intelligence.

To emphasize the significance of qualitative states in mental processing, it is helpful to refer to a classic framework that was used to focus study on different aspects of the mind. Traditionally, cognitive science categorized studies on mental processing into three groups: cognition, affection, and conation (Kant, 2023, p. 311), (McDougall, 1923, p. 266), (Hilgard, 1980), (Tallon, 1997). Cognition refers to the perceptual acquisition of knowledge through the senses, experiences and thoughts. Affection refers to experiential emotional states like sentiments, mood and attitudes. Conation refers to the connection between knowledge and behaviour. It encompasses the processes that bridge cognition and action. Qualitative experiences are a prominent feature of all these processes. They are leveraged towards perceiving, and understanding the world (cognition), adopting an appropriate attitude to the world (affection), and informed by cognition and affection, processing an appropriate response to the world in the form of behaviour (conation).

It is worth noting that this tripartite framework is not meant to suggest that these are separate, independent processes within the mind. Rather, it serves as a conceptual tool to focus attention on specific aspects of mental activity. As research has shifted from a holistic view of the mind to more granular exploration, the use of this framework has declined. However, I use it here because it provides a clear, intuitive way for a lay reader to understand how central experiential states appear to be in realizing the autonomy of cognitive processes.

Cognitive agents' ability to navigate complex environments and the dynamic problems they pose is the primary reason that we think of cognitive agents as generally intelligent. Cognitive agents are faced with a continuous stream of dynamic problems posed by nature in their native environments. In meeting these problems, the agents demonstrate cross domain problem solving, they demonstrate learning, and they do both things with an autonomy that makes it apparent that they are self-directed. The three lenses on the mind, cognition, affection and connation, highlight that experiential states appear to support the cognitive agent in the self-acquisition of information that defines problems, in self-motivating actions, and in operationalizing actions. All of which appear to be important for how the agents interacts and solves the problems in their environment. Experiential states appear to be important for the single example of general intelligence that we have to date.

It should be kept in mind that the type of general intelligence exhibited by cognitive agents is the inspiration for artificial general intelligence research. In a paper that constituted a pre-read for the Dartmouth Summer Research project where the AI research field was defined and launched, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon not only coined the term “artificial intelligence” but defined its ambition. They stated that AI research aimed to realize a machine that behaves “in ways that would be called intelligent if a human were so behaving (McCarthy, et al., 1955).” This first generation of AI researchers were labouring under ambitions that today would be classified in the emergent subset of AI research that seeks artificial general intelligence. Their stated ambition for AI research is almost indistinguishable from contemporary AGI researchers like Ben Goertzel and Cassio Pennachin who have defined the modern AGI research field as the body of research that aims to create machines with intellectual capability that emulates human-like general intelligence (Goertzel & Pennachin, 2007).

There is a general tendency in the AGI research field to adopt humans as the muse for the stated research agenda. This tendency is derived from and is sustained by an idea that general intelligence is best exemplified by humans. It is an anthropomorphic bias that probably stems from the fact that humans are more intimately aware of their own general intelligence than of any other form. It therefore makes sense to draw an understanding of intelligence and the inspiration of recreating it from what is known. However, as human study and knowledge about other types of cognitive agents have grown, humans have become familiarized with a broader range of examples of general intelligence in nature. A broad variety of cognitive agents confront dynamic challenges in varying environments with versatility and adaptability that matches, and in some instances betters human abilities. This has influenced contemporary AGI researchers to view the reference to human-like abilities in defining general intelligence as potentially limiting or even misleading. What is interesting about general intelligence for the AGI researcher is not limited to human general intelligence. The *sine qua non* of general intelligence, i.e., general problem solving that has adaptability, is robust, features innovative learning, reasoning, and is output in the form of behaviour, is common to other cognitive agents. Therefore, a definition of the AGI research agenda that is sanitised of anthropocentrism is that artificial general intelligence research aspires to create machines with intelligence that emulates cognitive intelligence.[2]

If the intelligence exhibited by cognitive agents is the inspiration for the artificial general intelligence research field, then why is a large part of the research field attempting to realize AGI models without a key feature of cognitive-supported intelligence, the role of experiential states? There is no occurrent instances of general intelligence that excludes the role of experiential states. Cognitive agents that exhibit general intelligence leverage process that are

[2] The thesis will maintain the more common anthropocentric reference to human intelligence in referring to the motivating influence of AGI research. The reader should recognize this as an attempt for uniformity with the canon and should read any such reference as interchangeable with cognitive intelligence.

supported by experiential qualitative states in forming knowledge of the world, adopting attitudes to the world, and processing responses to general problems in the world. It might be the case that it is possible to realize general intelligence without consideration for the role of experiential states. However, such an endeavour must be accepted and approached as a metaphysical leap of faith. The AGI engineer who pays no regard to the functional properties of qualitative states in general intelligence is gambling on the fact that general intelligence without the functional roles of qualitative states is a physical possibility even though all the examples of general intelligence that are known feature qualitative states prominently.

It is far more parsimonious to proceed under the impression that experiential states have important functional roles to play in general intelligence. These roles should be identified and duly accounted for in models of general intelligence.

1.3. Research agenda.

1.3.1. The Role of qualitative states in supporting autonomy in general intelligence.

In communicating that Tik-Tok did not have feelings, what Baum was communicating was that the contraption he envisioned did not leverage any form of experiential states in the role of feelings. We can be certain that Tik-Tok did not have phenomenal experiences. There is nothing to suggest that Tik-Tok had the underlying biochemical substrate that has been linked with accessing phenomenal attributes as we know them today. And, if we are to be charitable to Baum's prowess in effective articulation, we can also be sure that Tik-Tok did not access states that marked out feelings like joy, sadness, or sensations like pain or hunger. Tik-Tok somehow went about solving general problems in its environment without recourse to qualitative states.

The lesson that Tik-Tok appears to embody is that qualitative states have no role in general intelligence. This appears to be at odds with the only example we currently have of general intelligence, i.e., cognitive general intelligence.

As highlighted earlier, the cognitive general intelligence as instantiated by a mind features qualitative states very prominently. For humans, and probably other animals with close homological relations to humans, these qualitative states take the form of phenomenal experiences.

A *prima facie* case can be made that in cognitive intelligence, qualitative states in the form of phenomenal experiences play a role in the self-control that helps the cognitive agents navigate the complexities and dynamism of their native environment. These experiential qualitative states are active in cognition. For instance, consider how information from the external world reaches the body as perceptual stimuli. This information is experienced as phenomenal sensations that meaningfully reflect or refer to the external world. The qualitative aspect of perceptual experience appears to be crucial for providing information about the world and the problems it poses. Qualitative states also appear to play a central role in affection. For example, consider bodily sensations like pain, itches, and hunger; passions such as lust, anger, and fear; or emotional moods like elation, depression, and boredom. Intuitively, these phenomenal experiences seem to be involved in processes that motivate or inform action Qualitative states also appear to play a significant role in operationalising actions. Consider the impact of hedonic states. These are qualitative states that afford feelings of pleasure or displeasure. These hedonic states guide cognitive agents in making choices by allowing them to prefer the quality of one experience over another. This process of learning through preferences of phenomenal experience is essential for shaping behaviour.

In cognitive agents, qualitative states appear to be crucial for enabling self-direction in response to problems in complex and dynamic environment. To illustrate how these states are active in processing of behaviour, consider this scenario:

Person A feels hungry. This is a phenomenal experience that informs them of an internal state (cognition). This hunger also influences their attitude toward the environment and motivates them to seek food (affection). When Person A sees an apple on the table, the perception of the apple is a phenomenal experience that provides information about the environment’s affordances (cognition) and influences their attitude toward the affordance (affection). Motivated by hunger and previous positive experiences with apples (conation), Person A decides to eat the apple. In this example, Person A’s decision to eat the apple is meaningfully supported by qualitative states. It’s hard to imagine that Person A would have made the same decision if they hadn’t felt hungry, perceived the apple within reach, or had learned from previous pleasurable experiences with apples.

1.3.2. Computation and its lack of qualitative states.

The main approach to building artefactual models of intelligence is computation. Computing is the process through which a system uses algorithmic processes to simulate, describe, or project the transformation of information that represents states of affair or knowledge from the world. The input of a computing system is structural data that describes a feature in the world at T1. The structural data is input into the system as discrete, transduced and represented information that is manipulatable by the system. The process of computing then systematically manipulates the represented structural information under appropriate rules to realize a form that either simulates changes, describes, or projects the transformation in structural features of the input at T2. Computing is agnostic about processes. To date, there have been various approaches to building computational models of intelligence. The main approaches include discrete symbol processing, connectionism, statistical modelling, and hybrid approaches that combine some of these methods.

It is not inconceivable to think that computational models can be bult that are programmed to seek out apples, retrieve the apples and perhaps even process the apples to extract energy. These models could easily be integrated into a Robot A. Such as system, Robot A, will have some of the nous of navigating and engaging with the environment as Person A. If Robot A is built with contemporary approaches to computation, it will likely not have qualitative states that inform its processing.

Yet, regardless of how well Robot A performs in producing output like Person A, it will not satisfy our intuitions about it having intelligence that is comparable to cognitive general intelligence. This is hardly surprising. Afterall if we believed that Robot A satisfied our intuitions about general intelligence, AGI would not be an open field of inquiry. AGI seeks artefacts that possesses the ability to understand, learn, and apply knowledge like a human. Something about Robot A does not appear to satisfy us that it has human-like intelligence.

Robot A appears to be clever in the sense of being a complex contraption that performs programmed tasks appropriately. In which case, it is not clever in the sense that humans are clever. A significant amount of Robot A’s problem solving is credited to its designers. The designer has grounded the meaning that Robot A uses to define the problem and has prescribed the logic Robot A uses to solve the problem.

This is significantly different from how Person A approached the apple problem. They perceived the problem, were intrinsically motivated to solve the problem, and processed the solution without recourse to external agency and through processes that are wholly creditable to themself.

The Robot is doing programmed actions, while the Person is doing intentional actions. Orseau, et al. give voice to this difference by suggesting that we adopt the term ‘device’ for systems that are described as performing input-output mapping, and the term ‘agents’ to refer to systems that have intentional stances within computational theory (Orseau, et al., 2018).

Where did the two systems differ in their processes? Both Person A and Robot A appear to be doing operations that can be described as computation. They are processing represented information that describes the world by systematically manipulating it to realize an output in the form of action.

The obvious differences between the two appear to be the roles where qualitative states are writ large in cognition. Person A formed knowledge of the problem with the help of perceptual experiences (perceiving the world). Meaning that defines the problem in Robot A was prescribed by an external agent. Person A was motivated to solve the problem by an affective internal state (feeling hunger). Robot A is labouring under a prescribed motivation. It does not have its own experiential states. Person A operationalized their action with the help of hedonic feelings. Robot B is labouring under pre-defined or fitted instructions.

One could conclude that a significant difference between the two is that while both appear to be doing processes that can be defined by computation, Person A is computationally self-sufficient, while the Robot A is reliant on outside agency to effect and complete its computation. This difference in the type of autonomy in processing actions that the two exhibit is one of the reasons that we are reluctant to recognize Robot A as having intelligence that is like the Person A, even when they have matching input to output streams.

If cognitive general intelligence is the muse for AGI's goal of realizing artificial intelligence that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks, then it is arguable that AGI should aim to realize the computational self-sufficiency of human intelligence. The AGI artefact should be self-directed in recognizing problems in its environment and solving them. My assertion is that replicating and modelling the functional roles of qualitative states in the mind is key for accounting for computational self-sufficiency.

1.3.3. Centring the role of qualitative states in AGI research.

Although there is an intuitive and compelling understanding for the roles of qualitative states in cognition, these roles are overlooked by the technical side of AGI research, even when they recognize that they are labouring under ambitions to realize intelligence that is like cognitive intelligence.

The technical side, largely composed of engineers, focuses on building precise, reliable, and carefully calibrated products. Engineers prefer to base their work on principles supported by strict, demonstrable, and replicable axioms. The form of qualitative states in the mind, i.e. phenomenal experiences, has so far resisted such rigorous definition. Phenomenal experiences are perceived more as a subject of metaphysics than engineering. To many engineers, phenomenal experiencing is too open to diverse theoretical and methodological interpretations to serve as a solid foundation for engineering. Consequently, those working on practical AGI development often view the subject as outside their remit or only of incidental interest. However, if the goal is to achieve cognitive-like general intelligence, it would be a mistake to overlook the importance of qualitative states in supporting the self-sufficiency of the mind.

The good news for AGI engineers is that when considering the role of qualitative states in general intelligence, they do not need to solve the seemingly intractable mysteries that abound in conceptualizing phenomenal experiencing. The field of artificial general intelligence is primarily a functionalist pursuit. This means that the agenda is to create models of general intelligence that functionally resemble human (or cognitive) intelligence. The goal is not a faithful replication of the brain, or the biochemical substrates associated with human intelligence. While there may be a research agenda focused on recreating intelligence by replicating the brain's biochemical structures, this is so far removed from current AGI research that it would likely fall into a different research category.

Instead, AGI researchers seek to develop artefacts that exhibit general intelligence, without needing to reconstruct the specific organic structures that inspire the field. AGI engineers, therefore, must believe that general intelligence can be achieved through various materials and processes, that are defined by how well they function in producing output that comports to general intelligence, rather than by recreating the specific constitution of whichever intelligent system has inspired the project. As functionalists, their interest in qualitative states should lie in understanding the functional role these states play in cognitive general intelligence.

AGI researchers should set aside concerns about the specific form of qualitative states and focus instead on the functional roles these states fulfil in general intelligence. This focus can be separated from debates about the ontic nature of these states, much like how one can understand the effects of gravity without needing to grasp its essential nature.

The focus for the AGI researcher should be on replicating the causal effects of qualitative states in cognitive general intelligence. For example, scientists can simulate the effects of zero gravity on the human body using a zero-gravity plane or they can simulate the effect of increased gravity with a G-force centrifuge. Neither of these processes necessitate replicating gravity in its essential nature. It only requires the replication of the causal effects of gravity in some narrow but useful aspect.

Similarly, AGI research should concentrate on the functional properties of qualitative states in the narrow aspect of processing cognitive general intelligence. For this research, that interest is limited to the functional roles of qualitative states in supporting computational self-sufficiency. This interest should focus on resolving whether these functional roles are replicable in artificial models of general intelligence.

One approach AGI researchers can take to replicate the functional roles of qualitative states in supporting computational self-sufficiency is to explore whether these roles are uniquely realizable by their phenomenal manifestation. In philosophy of mind, the term "qualitative states" is often used interchangeably with "phenomenal states." A qualitative state is defined and accessible through its qualities or attributes, these qualities or attributes are phenomenal in character. The question for AGI researchers is whether the functional roles they aim to replicate in artificial general intelligence are narrowly dependent on these phenomenal manifestations, or can they be realized through some broader properties of qualitative states.

1.3.4. Separating qualitative states from their properties.

In humans, experiential states have qualitative form that bear certain properties. Qualitative states subsume sensory experiences like the redness one perceives when looking at a ripe tomato or the sweet taste of a ripe pineapple. It can also refer to the experiential aspects of emotions, such as joy or sadness, or the sensations associated with internal interoceptive perceptions, like pain or hunger. Qualitative states are defined and accessed by their distinct attributes, which, in humans, are inherently phenomenal in nature. In the words of Nagel, there is something that it is like to undergo the state (Nagel, 1974). Experiential states like the experience from eating a pineapple and states like joy or pain are qualitative states in the sense that they are accessible and defined by virtue of their distinctive quality. This accessible quality or attribute is phenomenal in character.

Certain properties characterize a qualitative state. As shown in the examples above, qualitative states, in the form of phenomenal experiences, can take various forms. For instance, the visual experience of seeing a red apple differs from the gustatory experience of tasting a pineapple. These, in turn, differ from the experiential aspects of emotions such as joy or sadness, or from interoceptive perceptions like pain or hunger. However, there are certain common properties that define all of them as qualitative experiences.

One such property is discernibility. Qualitative states are discernible. The defining attributes or qualities of a qualitative state are identifiable and accessible. For example, there is something distinct and accessible about seeing a red apple, feeling joy, or experiencing hunger.

Another property is differentiability. Take the examples cited above, the experiences of redness one perceives when looking at a ripe tomato, the experiential aspects of emotions, such as joy or sadness, or the sensations associated with internal interoceptive perceptions, like pain or hunger. All these states are phenomenal in character. However, they can be differentiated. Not only are they distinguishable from each other, but they can also be differentiated from having no experience. Qualitative states have the property of differentiability.

In addition to being discernible and differentiable, qualitative states are experienced as continuous. Continuity here can be understood as a stream or a succession of "specious presents," connected by smooth and consistent transitions over time (Levanon, 2016), (Thomas, 2023). The qualitative states mentioned above, such as redness, sweetness, joy, sadness, hunger, or pain, are all experienced as continuous streams. Even when qualitative features have quantifiable elements, these elements are presented as part of an ongoing present.

Finally, qualitative states are potentially variable. This means that their defining qualities or attributes can change or vary in intensity. For example, the redness of an apple, the sweetness of a pineapple, the feeling of sadness, or the feeling of hunger can all be experienced in varying degrees of intensity.

Therefore, at least four key qualitative properties characterize qualitative states:

1. **Discernibility:** They can be recognized or identified.
2. **Differentiability:** They can be distinguished from other qualitative states and from nothingness.
3. **Continuity:** They are experienced as a continuous stream of moments.
4. **Variability:** They have the potential to vary or change in intensity.

Why bother with defining qualitative properties? One reason is that we want to address the topic of qualitative states holistically, so it is sensible to typify them by what they have in common, i.e. qualitative properties, rather than by their phenomenal attributes which come in an unwieldly number of types (e.g. perceptual, emotional, affective), and expressions (e.g. the different types of perceptual, emotional, and affective experiences).

Secondly, defining experiential qualitative states by their phenomenal character risks accusations of anthropocentrism or related biases. We have access to phenomenal experiences through first-person encounters. In that sense, I am certain through an introspective cartesian method that my experiential states have phenomenal attributes because I have a first-person experience of ‘something it is like’ to have an experiential state. I can also reasonably infer that other humans have experiential states with phenomenal attributes. However, this certainty becomes less and less justifiable the further I cast my net from things that are like me. Mammals are close enough to humans to suspect that they have phenomenal experiences. This might be expanded to all animals, but not without a great degree of controversy depending on your outlook. How about plants, fungi, protist, monera or aliens, do they have experiential states with phenomenal character in the same sense?

Take for example plants. Convincing arguments have been made that they don't have phenomenal experiences (Hamilton & McBrayer, 2020), (Mallatt, et al., 2021). However, it is not controversial to say that they sense and respond to natural stimuli. In some cases it is undeniable that plants use the sensed stimulus to process complex responses that have been described as cognitive intelligent behaviour (Brenner, et al., 2006), (Leopald, 2014), (Trewavas, 2016). It is one thing to say that plants that sense and respond to these natural stimuli have states with qualitative properties, and it is another to say that they have qualitative states that are phenomenal in character. The former is more considered. If for nothing else, it is worth distinguishing between qualitative properties and qualitative states in the form of phenomenal experiences as a matter of circumspection.

AGI researchers should investigate whether these qualitative properties, independent of the phenomenal aspects, can have causal efficacy in models of general intelligence.

It can be difficult to conceptualize qualitative properties without relying on the phenomenal forms with which they are made manifest in the mind.

As an exercise, it might help in this cause to think of the phrase qualitative state *ad litteram*. The task here is to bracket the use of the phrase 'qualitative state' as it is used in in the philosophy of mind where it is intractably linked to phenomenal attributes or qualities, and to think about it in its most literal sense. A qualitative state in the most literal sense is a physical state defined and made accessible by a defining quality or attribute. While this quality or attribute can be phenomenal as in the case of phenomenal states in the mind, in this broader reading, we can imagine other physical systems (other than the mind) with other natural defining qualities or attributes (other than phenomenal attributes). For example, natural attributes such as luminosity, sound, pressure, redolence, or motion. A physical system with states defined by these attributes can be in an *ad litteram* reading defined as being in a qualitative state. For example, an overheated laptop (a physical system) is in a qualitative state (*ad litteram)* of high heat. That is to say that it is defined by an accessible thermic quality.

A qualitative state described in this *ad litteram* reading of the phrase, will have the same qualitative properties as the qualitative states defined in the philosophy of mind. An *ad litteram* qualitative state is a state in a physical system defined and made accessible by natural qualities like luminosity, sound, pressure, redolence. These states in a system will have objective qualitative properties beyond the phenomenal forms in which humans perceive them. These states are distinct and identifiable, making them discernible and differentiable from one another or from nothingness. They are continuous, meaning that the qualities like light, sound, motion, and pressure that define them present as ongoing streams of specious present moments. Even when these qualities have quantifiable elements, they are presented as part of a continuous present. Lastly, these states are potentially variable, meaning that the defining qualities or attributes can vary in intensity. For example, the brightness of light, the amplitude of sound, or the force exerted by pressure can all vary.

The point of this exercise; taking an *ad litteram* definition of qualitative states, is to highlight that qualitative properties are not the reserve of the phenomenal manifestation that defines the phenomenal qualitive state.

The question the AGI researcher should seek to resolve is whether these qualitative properties i.e., distinctness, discernibility, continuousness and variability are the efficacious attributes of the qualitative states in cognitive general intelligence or whether the efficaciousness of qualitative states in cognitive intelligence is wholly dependent on the phenomenal form by which they are made manifest in the mind.

There is substantial evidence to suggest that the efficacy of qualitative states in mental processes can be attributed to their qualitative properties such as discernibility, differentiability, continuity, and variability, rather than the specific phenomenal forms they assume in the mind. Research indicating cognitive intelligence in life forms that do not appear to possess phenomenal attributes supports this view. For instance, studies on plant behaviour suggest that plants exhibit cognition-based intelligence, despite lacking clear evidence of phenomenal experiences (Hamilton & McBrayer, 2020), (Mallatt, et al., 2021).

Additionally, studies have shown that phenomenal experiences can sometimes be interchangeable in the execution of certain cognitive functions. For example, it has been demonstrated that blind individuals can learn to 'see' through tactile stimulation (Chekhchoukh, et al., 2013), (Friberg, et al., 2011) or auditory signals (Ward & Meijer, 2010), (Kolarik, et al., 2017), instead of the visual signals typically associated with sight. These experiments in sensory substitution indicate that the efficaciousness of qualitative states in cognition is due to the shared qualitative properties such as discernibility, distinctness, continuity, and variability rather than the specific phenomenal form of a qualitative state.

Thus, researchers in artificial general intelligence (AGI) should feel reasonably confident in focusing their attention on seeking the replicable roles of qualitative properties, in cognition rather than focusing on replicating phenomenal forms,

1.3.5. The research question.

This research aims to answer the question ‘How can we replicate the functional roles of qualitative properties in cognitive processes to realize artefacts that have computational self-sufficiency that is functionally comparable to the self-sufficiency of cognitive general intelligence?

Folk psychology tells us qualitative states in the form of phenomenal experiencing have at least two key roles in the form of computational self-sufficiency that is exhibited by cognitive general intelligence. (1) Qualitative states in the form of phenomenal experiences facilitate the acquisition of knowledge from and of the world and the problems it poses. (2) Secondly, qualitative states in the form of phenomenal experiencing have roles in the ratiocination that yields intentional actions. What is needed is a physical (naturalist) account of these processes.

A central assertion of this research is that the fundamental efficacy of qualitative states in supporting computational self-sufficiency is traceable to qualitative properties rather than the specific phenomenal forms of these states. The qualitative properties of discernability, differentiability, continuousness and variability are the key efficacious properties of qualitative states in cognition. Therefore, it is essential to delineate how these qualitative properties fit into the physical causal chains that underlie the two pivotal roles for cognitive computational self-sufficiency: grounding understanding and directing self-creditable decision-making.

To address the main research question, this thesis will explore how qualitative properties contribute to the introducing of meaning from the world into mental processes and are efficacious in propositional states that inform self-controlled decision-making in cognitive processes. Based on these explorations, the thesis will identify two replicable principles: (1) the Analog Principle and (2) the Internal Model Principle. These principles define the functional roles of qualitative properties in grounding knowledge and processing self-controlled decision-making within cognitive general intelligence. Furthermore, the thesis will examine whether models built with these principles can be attributed with the same intentional actions that characterizes cognitive general intelligence.

It is important to emphasize that any AGI model(s) described in this thesis are primarily intended to demonstrate that qualitative properties have replicable functional roles in realizing computational self-sufficiency that is comparable to cognitive intelligence. To that end it is necessary only to define a minimal viable product of artificial general intelligence. A minimally viable product or MVP is a concept borrowed from product development. An MVP is the most basic version of a product that includes only the essential features necessary to address a core problem or need. Examples proffered as minimally viable AGIs in this thesis should not be interpreted as a claim that the problem of AGI has been solved. They should be understood as demonstrations of essential features. In this case the essential feature I aim to demonstrate is computational self-sufficiency. However, considering that technological advancements often correlate with increased understanding of principles and technical expertise, it is reasonable to suggest that the principles outlined here could potentially be useful in more advanced AGI research.

1.4. Scope and key descriptions.

1.4.1. Scope.

Given that AGI research is closely related to understanding the mind, it naturally overlaps with many theories and discussions in the philosophy of mind. However, it would be unwieldly in this thesis to refer to all the relevant work in the philosophy of mind canon. To maintain narrative coherence, this thesis selectively navigates key topics in the philosophy of mind. For example, it adopts an implementationist computational theory of mind as a foundational assumption, with only space for a brief justification for this choice in Chapter Three. An implementationist computational view does not recognize a conflict between Turing style symbol processing and connectionism. It recognizes them as different levels of implementation or different ways of performing the computations of the mind (Marcus, 2001), (Smolensky, 1988). Chapter Four introduces the Analog Principle, which is a causal theory of mental representation. While Chapter Five discusses the Internal Model Principle which aligns with mental model theories. Chapter Six examines self-directedness through an instrumentalist intentional lens that draws from Daniel Dennett's intentional stance (Dennet, 1987, p. 17). For each of these choices, there is a significant body of philosophical work that could be cited to support or critique the chosen positions. Such depth would be unwieldy within these pages.

While the narrowness of the path cut through the canon is mainly due to word limit, expediency, and in support of narrative clarity, there is also a method to selecting the path. AGI research has increasingly taken on a scientific bent influenced by its practical branch that is led by engineers. The path cut through the canon is responsive to the scientific practice of the research field. The choices made in selecting a path through the canon has favoured theories, principles and definitions that aim to respond to, build on, or improve on scientific practice in the research field that has shown promise. There is a recognition here, that while the practical end of the research field has not yet yielded AGI, it has registered progress worth building upon.

I recognize that there are certainly other justified choices that can be made in curving a path through the philosophy of mind canon to support an account of AGI. I have left that work for others. I also recognize that even in the choices I have made, depth was sacrificed for breadth. For example, Chapter Four has a discussion on Searle's Chinese Room that had to omit a discussion on the vast retorts it has attracted. Chapter Five has a brief discussion on the sense of body ownership as it relates to the minimal self which hints at work in 4E Cognition theories. However, there is insufficient space here to explore such linkages. I believe that the strength of this thesis lay in telling the whole story.

With this disclaimer registered, below are brief definitions of several key terms central to this research. More detailed explication and defences for these descriptions are included as they arise in the unfolding narrative.

1.4.2. Key descriptions.

Chapter Two lays out a discussion of the type of artificial general intelligence that is of interest in this thesis. This discussion includes a dissection of the key terms, 'intelligence', 'generality' and 'functionality' that are associated with AGI research. However, it is worth describing here in brevity the definition of intelligence adopted and how it applies to narrow and general intelligence.

The word 'intelligence' originates from the Latin words *intelligentia* or *intellēctus,* which are both nouns derived from the verb *intelligere*, which means to comprehend or perceive. In modern usage the word has been defined in several ways that have attracted a great deal of debate and controversy (Legg & Hutter, 2007). The controversy mainly centres on what abilities qualify for intelligence, whether these abilities can be quantified, and who or what possesses these defining abilities. The abilities often associated with intelligence include the capacity for rationality, abstraction, logic, understanding, self-awareness, creativity, and critical thinking.

It should be kept in mind that AI theory has a practical bent. The aim is to realize an engineered artefact that can be put to use. The definition of intelligence that is interesting to AI theory is a definition that captures what intelligence affords. Of what use are all these capacities that are the subject of agitation in identifying the definitive quality of intelligence? What does having the capacities for rationality, abstraction, critical thinking, creativity, and understanding afford? AGI will benefit from a functional view on intelligence.

Humans and animals that possess one or more of these capacities like rationality, abstraction, critical thinking etc, leverage them to solve problems in their environment. AI theory seeks to build machines that have intelligence that is suitably comparable to humans or animals. The research field seeks capacities that can be leveraged for problem solving in a manner that is suitably comparable to animals or humans. In simple terms, the definition of intelligence that is adopted in this research is that intelligence is the capacity for problem solving.

Narrow intelligence is the singular capacity to solve a domain restricted problem. References to narrow intelligence in this research will pertain exclusively to artificial narrow intelligence. Artificial narrow intelligence refers to the capacity of an artefact to have a singular function that solves a domain restricted problem. The term is also used to refer to an artefact that possesses narrow intelligence. To differentiate between the two uses of the word the paper will use the long form ‘artificial narrow intelligence’ to refer to the capacity to solve a single domain restricted problem, and the short form ‘ANI’ to refer to an artefact with narrow intelligence.

General intelligence is the capacity to solve both novel and recurring problems in a native environment. General intelligence is not restricted to a single domain like narrow intelligence. However, this does not mean that general intelligence refers to unrestricted problem-solving capacity. There will undoubtedly always be problem-solving constraints occasioned by the physical limitations of a system with general intelligence. Because of this, general intelligence tends to be optimized for a native environment. A native environment is an environment in which the subject with general intelligence has been optimized, either through adaptation or configuration to inhabit or to address problems. Two things are essential for locating general intelligence. The first is that there must be cross-domain problem solving capacity (i.e., the ability to solve more than one problem). The second is an ability to learn to solve unfamiliar problems or differently contextualized problems in the native environment. Artificial general intelligence refers to the capacity of an artefact to exhibit general intelligence. It also refers to an artefact that possess general intelligence. Like in the case for narrow intelligence, the paper will use the long form, ‘artificial general intelligence’ to refer to the capacity, and the short form ‘AGI’ to refer to the artefact with general intelligence.

Throughout this paper, the term "qualitative" will be frequently referenced, primarily in line with its common use in the philosophy of mind. Here, "qualitative" refers to the phenomenal attributes of experiencing. A qualitative state in this reading is therefore a state in mental processes that is defined by a distinct quality or attribute which in this case is phenomenal in character.

The phrase ‘qualitative properties’ is used to reference the properties of *qualitativeness*. Qualitative states in the mind have phenomenal attributes that possess the properties of discernibility, differentiability, continuity and variability. These qualitative properties can also be conceptualized independently of the phenomenal attributes typically associated with qualitative states in the mind. To illustrate this, consider an *ad litteram* interpretation of the term "qualitative state". In this reading a qualitative state is very literally a state defined by distinct natural qualities or attributes. The defining quality or attribute need not be phenomenal in nature; it could be a natural attribute such as luminosity, temperature, or redolence. A physical system in a state defined by such an attribute would still exhibit the qualitative properties of discernibility, differentiability, continuity, and variability. Whenever this literal interpretation of a non-phenomenal qualitative state is referenced, it will be explicitly qualified with the phrase "*ad litteram*."

Computation has traditionally been the foundation of AI theory. It refers to the process where a system uses algorithms to simulate, describe, or predict transformations of information that represent real-world states or knowledge. The system's input consists of structured data that describes a feature of the world at a given time (T1). This structured data is fed into the system as discrete, processed, and represented information that the system can manipulate. Through computation, the system systematically processes this information according to specific rules, producing an output that either simulates changes, describes, or predicts transformations in the structural features of the input at a later time (T2).

The computational theory of mind posits that human and animal minds share enough characteristics with computation to support the idea that the mind performs computations. The theory can take various forms, depending on how computation itself is understood. Typically, computation is conceived in terms of Turing machines, which process symbols based on a set of rules and the machine's internal state. A key feature of the CTM is that it is agnostic about the physical specifics of the machine performing the computation. For example, computations could be performed by binary switches or biological neural networks, as long as the system manipulates represented inputs and internal states according to rules to produce appropriate outputs. CTM asserts that the mind is not merely comparable to a computer program but actually operates as a computational system. Supporters of this perspective are often referred to as computational cognitivists or computationlaist.

The thesis will make frequent references to goal-directed actions. The thesis is interested in a very particular type of goal-directedness. It is interested in intentional actions in its most basic conceptualisation as it relates to action theory. An intentional action is a behaviour driven by an intention, or an expectation that the behaviour is likely to bring about a desired outcome (Dickinson, 1985). Leonard Dung suggests that agents with intentional agency can be identified by autonomy as well as the ability for directly impacting the surrounding world, long-term planning and acting for reasons (Dung, 2024). Dung’s definition of intentional agency would subsume the goal pursuit that is exhibited by all cognitive agents. This is interesting to this thesis because AGI research aspires to general problem solving that is functionally comparable to cognitive general intelligence. The thesis will take Dennet’s instrumentalist view in which attribution of intentions to a system is dependent on the usefulness of the system’s attitude to bodily and environmental states to determine, explain and predict the system’s behaviour. This contrasts with the realist views that locate the intentions that inform goal-directedness in a presumed unique feature of organic mental states (Dennet, 1987). The thesis’ interest in intentional agency is limited to resolving the question of whether the AI models it describes are goal-directed in a way that is analogous to cognitive agents. It will not pronounce itself on debates in AI agency that aim to address questions of morality or locating accountability and culpability in AI systems. These are important questions, but they lie just outside the scope of this thesis.

1.5. Layout.

The first chapter has introduced the motivation for the thesis, the central claim in the thesis, the research agenda, and the tone and layout of the thesis. The thesis is motivated by the anticipation of the social and economic value of realizing artefacts with artificial general intelligence. If developed with due regard to ethical considerations, AGIs have the potential to positively improve human lives. The central claim underpinning the research is that the artificial general intelligence research field has ignored a potentially key desideratum for general intelligence in the form that it is known today. The mind is computationally self-sufficient. If the mind is the muse for artificial general intelligence, then AGI R&D will do well to take an interest in engineering for the functional roles that support the mind's self-sufficiency. This thesis will aim to articulate the physical principles that describe the functional roles of qualitative experiencing in the processes that yield computational self-sufficiency in cognitive general intelligence.

Chapter Two will define and defend a concept of general intelligence that is used in this thesis. Currently, AGI research serves a dual purpose as both a scientific exploration and a practical engineering discipline. The discovery aspect focuses on identifying the principles underlying general intelligence, which can then be used to inform engineering efforts to build AGI. This dual approach is atypical in engineering, which usually builds on established scientific foundations. Since AI theory is still in discovery, AGI research remains open to varied definitions and approaches. This chapter aims to clarify and justify the specific concept of AGI that guides this thesis. Defining the form of general intelligence relevant to AI theory is crucial, as it sets clear goalpost for the research field. The definition proposed and defended here will outline the qualities an artefact must possess to be considered an AGI in what is termed the 'Grand Ambition.' The Grand Ambition is to create artefacts with general intelligence that is functionally comparable to human intelligence.

Chapter Three explores why popular computational approaches to artificial general intelligence fall short of achieving the Grand Ambition. It evaluates computational models that have been proposed as examples of AGI and compare them against the necessary and minimally sufficient properties derived for the Grand Ambition. The chapter will demonstrate that these models fall short due to a fundamental conceptual flaw. Given the inadequacy of these computational approaches, the chapter will investigate whether computation is indeed a suitable path to achieving general intelligence that is comparable to human general intelligence. While it will show that the mind shares some characteristics with computational processes, it will also argue that the mind goes beyond simple input-output mapping. The mind is self-controlled in a manner that makes it self-creditable for its computation. The chapter will identify two key issues (i) the grounding problem and (ii) the self-control problem, that prevent computational approaches from fully replicating the self-sufficiency of the mind. Finally, the chapter will propose and defend the replication of how these problems are solved in the mind, as a path to develop computational models that are closer to achieving the goals of the Grand Ambition.

Chapter Four focuses on the Grounding problem. It explores how the mind solves the grounding problem by self-grounding cognitive items with underived meaning directly from the world. Computational models of intelligence tend to overlook this self-acquisition of meaning, instead depending on external agents, like programmers or designers, to assign meaning to computational items. This reliance on external grounding has fuelled critiques from anti-cognitivists, who use this reliance on extrinsic agency to argue against the mind as a computer analogy. The chapter introduces the concept of symbol grounding and presents the Analog Principle (AP) as a key mechanism for self-grounding mental items. The principle offers a natural explanation for a causal theory of mental content. To demonstrate the veracity of the Analog Principle, the chapter applies the principle to counter Searle's Chinese Room argument. The chapter concludes by emphasizing the Analog Principle's crucial role in advancing the type of computational self-grounding that would satisfy the AGI's Grand Ambition.

Chapter Five explores how the mind manages to effect self-control. Current computational approaches typically depend on external agents that both define the problems to be solved and ground computational items with meaning related to the problem. This dependence on extrinsic agency gives the external agents most of the credit for control in computational problem-solving. However, the mind does not rely on external agents for self-control. It operates autonomously. Intuitively, the mind's self-control suggests the presence of an internal entity, often understood as an 'inner self,' which experiences the world through sensations and initiates actions. This chapter argues that the perceived inner self, which is credited with self-control, arises naturally from qualitative experiencing. It also posits that this inner self functions in a manner that is similar to an internal model in the Internal Model Principle (IMP). In engineering, the IMP describes an inner controller that enables a system to adaptively regulate itself in changing environments. By drawing on these functional similarities, the chapter proposes that the IMP can guide the design of computational systems capable of self-directed problem-solving in a manner that is functionally comparable to human intelligence.

Chapter Six presents a model developed according to the principles outlined in Chapter Four (the Analog Principle) and Chapter Five (the Internal Model Principle). Daniel Dennett argued in *Why Not the Whole Iguana?* that a physicalist theory is ultimately validated by modelling a complete cognitive organism (Dennett, 1978). Following this idea, Chapter Six introduces a basic, minimally viable model of computational self-sufficiency, built using the principles defined in the preceding chapters. The chapter then assesses whether this model demonstrates the intentional autonomy that would meet the requirements for artificial general intelligence. This evaluation is necessary because, while one might agree that the Internal Model Principle and Analog Principle support computational self-sufficiency, there could still be doubt as to whether this form of self-sufficiency enables intentional behaviour. Intentional actions are an essential component of the type of general intelligence that is like human intelligence. This chapter evaluates the model's states and functioning against philosophical arguments defining intentional action. It demonstrates that there is no reason to withhold intentional agency from the model.

Chapter Seven concludes by summarizing the thesis's main arguments and merits, as well as acknowledging its limitations. It also anticipates directions for future research. The philosophy of mind offers a wealth of resources, ideas, and approaches that could be highly beneficial to AI theory. This closing chapter states that AGI research can benefit from translating key insights from the philosophy of mind to advance AI theory.

1.6. Conclusion.

Tik-Tok of Oz perfectly captures the motivation for creating machines with human-like intelligence. It is easy to grasp how useful it is to have in your service a machine with intelligence on par with humans but with none of the feebleness that comes with the human body and will. This same motivation is driving artificial general intelligence research today.

Feelings are often considered optional or non-essential in computational models of intelligence. L. Frank Baum illustrated this view with his character Tik-Tok, who lacks feelings yet is depicted as having human-like cognition. Many in the AGI research field similarly dismiss the need for feelings in building human-like intelligence. However, in omitting feelings, Baum may have overlooked a key component of general intelligence. While fiction can sidestep this issue, AGI research might not. If feelings, in the form of qualitative states, play functional roles in cognitive general intelligence, then AGI research aimed at human-like intelligence will need to understand and replicate these roles in computational models.

An attempt to create AGI must begin with a clear end goal. Currently, there is no consensus on the precise features for AGI. To move forward, it is essential to define and justify the minimum viable properties that characterize the type of general intelligence of interest to this research. The next chapter will define and defend the necessary and sufficient properties for the type of general intelligence that interests this research.

Chapter 2: On the Mechanical Turk and the necessary and minimally sufficient properties for General Intelligence.

> *"That an AUTOMATON can be made to move the Chessmen properly, as a pugnacious player, in consequence of the preceding move of a stranger, who undertakes to play against it, is UTTERLY IMPOSSIBLE."*
>
> The speaking figure, and the automaton chess-player, exposed and detected. (Thicknesse, 1784).

2.0. The Mechanical Turk and its sufficiency as an AGI.

In 1770, Wolfgang von Kempelen constructed an automaton designed to play chess called the Mechanical Turk. The automaton that was intended to impress the Empress Maria Theresea of Austria, consisted of a mechanical man that sat in front of a chessboard. The automaton could survey the chessboard by moving its head from side to side, and it could move its hands and fingers to pick up and place chess pieces. The Mechanical Turk was able to play chess against experienced human players, perform the Knights Tour, a trick in which a knight is moved to occupy every square of the chessboard exactly once, and could communicate through a letter board with its human opponents on topics ranging from its age, marital status, and its inner workings, in English, French, and German.

Was the Mechanical Turk the first AGI artefact? Given that the automaton could play chess by physically moving the pieces, could perform the Knight's Tour, which while related to playing chess, aims at a different goal to a typical chess game, and could communicate in different languages on a range of topics, it is safe to say that the Mechanical Turk could apply problem solving across at least three different domains. The Mechanical Turk also exhibited flexible rationality. It responded appropriately to unpredictable human behaviour. This was apparent in how it could engage its human interlocutors in conversation on a range of different subjects at their whim and prompting, and the way it could identify different attempts at cheating and react to them appropriately. This flexibility of mind demonstrated that it could recognize and appropriately engage with new or differently contextualized problems in its native environment. If general intelligence is simply the capacity to engage in problem solving across different domains and the ability to address new or differently contextualized problems in a native environment, then by the letter of the law, the Mechanical Turk was an AGI artefact.

What made the Mechanical Turk a spectacle to behold was that it possessed the necessary output for general intelligence. It was capable of outputting behaviour that was on some level on par with human intelligence.

However, the literati of the day were hesitant to recognize the Mechanical Turk as an artefact that possessed general intelligence. Robert Willis stated in 1821, that the Mechanical Turk's capabilities could not possibly belong to a machine, they were "the province of intellect alone (Willis, 1821, p. 11)." Willis' position appears to suggest that there is more to general intelligence than the necessity to output actions that have the appearance of general intelligence. Willis hints at what is missing from the Mechanical Turk when he asks where the Mechanical Turk's intelligence is located, where can we find the "promethean heat" that animates the automaton (Willis, 1821, p. 15)? Willis appears to be hinting that to qualify as intelligent, there must be something within the mechanism that can be credited with generating the output that is deemed intelligence. It appeared inconceivable that just *any* mechanical whirring, even if it produced the right output, was sufficient to qualify as general intelligence.

The idea that the Mechanical Turk did not have the sufficient properties to be deemed an example of a general intelligence were further strengthened when it eventually transpired that the Mechanical Turk's 'promethean heat' lay in a human who was cleverly concealed within the mechanical guts of the machine. Willis' inquiry into where to locate the 'promethean heat' of the Mechanical Turk had intuitively hinted that, if it existed at all, it would be found within the Mechanical Turk. It did exist, and it was located within the Mechanical Turk. It was in the form of a chess-master squeezed inside the contraption. Yet, the discovery that the Mechanical Turk's intelligence was credited to something (in this case someone) within the Mechanical Turk, served to validate the naysayers rather than support the Mechanical Turk's claim to intelligence. This suggests a clarification to what is sufficient to qualify a contraption with the right output as generally intelligent. It would appear in this case, that for intelligence, it was not enough to credit something within the Mechanical Turk, but rather something that unambiguously belonged or was part of the Mechanical Turk itself.

Even with its secret laid bare, the Mechanical Turk raises some interesting questions. What are the definitive properties of general intelligence? How will one know when they are faced with general intelligence? The lesson from the Mechanical Turk appears to be that the extrinsic-facing properties of general intelligence, i.e., cross domain problem solving and the ability to solve new or differently contextualized problems in a native environment are necessary for general intelligence but are not sufficient properties for general intelligence. There is need to define what properties qualify as minimally sufficient for an artefact to be credited with general intelligence.

For ongoing exposition there is a need to identify clearly what this research deems an AGI in the minimal sense. There seems to be little resistance to the necessary characteristics of an AGI, which are cross domain problem solving and the ability to confront new and differently contextualized problems in a native environment. Yet little attention has been devoted to the sufficient properties that will qualify an AGI. The example of the Mechanical Turk demonstrates that this is important work. For among other reasons, it will establish a target and a means by which to gauge success in AGI R&D.

This chapter will identify and defend the minimally sufficient properties that, along with essential extrinsic-facing properties, such as the ability to solve cross-domain problems and tackle new or differently contextualized challenges in a native environment, would qualify an artefact as having general intelligence.

Section 2.1 will introduce a brief history of the AI research field and its quest towards what the thesis has titled the Grand Ambition of the research field. The Grand Ambition is to realize artefacts with general intelligence that is functionally comparable to human general intelligence. Section 2.2 will distil from the Grand Ambition, the necessary properties that will qualify an artefact as an AGI. The necessary properties are the extrinsic-facing or behavioural output of the artefact that most appropriately conform to the Grand Ambition. Section 2.3 will introduce and argue for the minimally sufficient properties that together with the necessary extrinsic properties will qualify an artefact as an AGI. Section 2.4. will defend the adopted definition of general intelligence. The last section, 2.5., will conclude by comparing The Mechanical Turk against both the necessary and sufficient properties to identify why it fell short of AGI status.

2.1. The Grand Ambition of AI research.

An official start date for the AI research field can be located at the Summer Research project on Artificial Intelligence held in Dartmouth in 1956. The research project lasted six to eight weeks. The central agenda for the project was brainstorming aimed at organising research around and about thinking machines into a coherent research project. Among other things, the pioneers who attended the conference settled on the name ‘Artificial Intelligence’ as a name for the research field. Speaking for a host of other pioneering luminaries who defined and officially launched the AI research field, John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon defined the ambitions of AI research as a goal to realize a machine that behaves “in ways that would be called intelligent if a human were so behaving (McCarthy, et al., 1955).”

The AI research field is in equal parts a field of discovery and an engineering discipline. The agenda of the research field as imagined by the Dartmouth group was to discover theoretical models and abstractions of systems that could explain or reproduce intelligence that is comparable to human intelligence, and subsequently build these models to realize an artificial intelligence artefact. Most of the attendees at Dartmouth would be recognized today as theoreticians rather than engineers. Their primary mandate was to discover theoretical models for systems that could explain intelligence. Over time, the field has evolved, shifting from predominantly theoretical research to a more practical approach led by engineers.

Since the Dartmouth conference, the AI research field has yielded unexpected insights regarding the nature of intelligence. AI research has revealed that intelligence can be characterized in at least two forms. Specifically, distinctions can be made between narrow intelligence and general intelligence. This revelation has emerged from the unfolding progress in the research field. Through progressive, and sometimes inspired development, the research field has realized incontestable success in reproducing intelligence across narrow domains, but has yet to match this success for general intelligence. It is unlikely that the pioneers of the research field would have anticipated that the demands for narrow intelligence would differ in a significant way to the demands for general intelligence. Nor would they have imagined how successful we would become at developing artificial narrow intelligence, and how intractable a problem artificial general intelligence would turn out to be. As a result, it is difficult to judge the research field for success against its original goals.

A section of AI researchers will look at the success achieved in narrow intelligence research and how ingrained it has become in society and argue that the research field has been successful. A machine is intelligent in the context of artificial narrow intelligence if, on the basis of information it is provided within a certain context, coupled with related information for the desired output, it can produce the desired output on new information within the same context (Aleksander & Hanna, 1976). This intelligence is like human intelligence in that it describes to some extent the way humans process certain narrow tasks such as classification, identification, pattern recognition, and the utilization of these capabilities in certain repetitive tasks. It is desirable because it produces outputs that are like human outputs or would be desired by humans in narrow domain tasks. Given that ANIs can leverage machine resilience, speed, and dedicated computing resources, they can often outperform humans in these tasks. These narrowly intelligent machines can be used as tools that are wielded by humans or incorporated into broader systems to solve narrow problems faster and more efficiently. Artificial narrow intelligence researchers might be compelled to argue that the forward-looking ambition of the research field should be to build more powerful (but safe) machines in this vein.

Another section of researcher might look at how the research field has failed to make any apparent progress towards general intelligence and argue that the research field has not delivered on its original goal. Their interpretation of the original goal was to build machines that are comparable to human intelligence in a broader and more direct sense rather than the limited comparability ANIs have realized in certain narrow tasks. Just like human intelligence, the machines sought by AI should be self-driven and have the capacity to acquire complex problem-solving efficacy through interaction with the environment or through being taught. AI research should aspire to general intelligence that is functionally comparable to human intelligence in this broader sense. This ambition is trumpeted by contemporary AGI researchers like Ben Goertzel and Cassio Pennachin who have defined the modern AGI research field as the body of research that aims to create machines with intellectual capability that emulates human-like general intelligence (Goertzel & Pennachin, 2007). Rather than be a mere narrow domain tool, a well-engineered AI artefact should be expected to identify tasks through engagement with the environment and implement these tasks on its own, or by working alongside humans in accomplishing tasks.

The primary interest of this research is to inform the ambitions of the latter group, the researchers who seek to realize an AGI artefact as an end goal of AI research. This is what this thesis will term **the Grand Ambition of AI research.** The Grand Ambition of the research field is to aspire to realize machines that have general intelligence that is functionally comparable to human general intelligence. It is arguable that the modern-day AGI researchers will find congruence with the pioneering class of researchers around the Grand Ambition. When the pioneering class set out the ambition of realizing human-like intelligence in a machine they would not have known that general intelligence was characteristically different from narrow intelligence. Their ambitions would have been informed by comparability to the broader definition of intelligence that humans expressed.

However, even within the Grand Ambition, there is room for disagreement about the agenda. The problem is twofold.

First, while the Grand Ambition states that the end goal is a machine that has intelligence that is functionally comparable to human general intelligence, it does not in certain terms identify what success looks like. Terms like 'intelligence,' 'general,' and 'functionally' are featured prominently in the Grand Ambition, but their precise interpretation is not specified. Without a consensus on the meaning of these terms, it becomes challenging to determine how the research field will recognize success in realizing an AGI. The example of the Mechanical Turk demonstrates that the research field needs a well-defined end goal. In the case of the Mechanical Turk, it appeared apparent that all the terms of the Grand Ambition had been satisfied. The Mechanical Turk exhibited general intelligence that was functionally comparable to human general intelligence. This remained true, even when its secret was exposed. If one considers the Mechanical Turk as system that subsumes its hidden player, it could not be contested that that its intelligence was humanlike. Yet no one would state that the Mechanical Turk was an AGI. The Grand Ambition does not satisfactorily communicate a target for the research field. There is a need to establish a shared vision on not only the necessary properties of an AGI, but the sufficient properties as well.

Secondly, there are no agreed upon methodologies, approaches or principles that will yield the Grand Ambition. The research field has remained accommodative of various approaches, with the view that this type of openness will increase the chances of stumbling on success. One might argue that this trial-and-error approach is a consequence of the research field sidelining theoreticians and leading with experimentation. The example of the Mechanical Turk demonstrates how understanding the functioning of a general intelligence is essential in confirming its status as an AGI. The learned folks at the time were adamant that they had to discover from where the Mechanical Turk generated its 'promethean heat' before they could be satisfied that it was an AGI. In seeking to investigate the Mechanical Turk's inner workings, the investigators were seeking functioning that could sufficiently be credited with general intelligence. It is doubtful that they had a specific configuration in mind for confirming intelligence. It is more likely that they had a sense that they would intuitively recognize processes that were sufficient to yield intelligence. Or at least recognize processes that could not possibly yield intelligence. Indeed, when the Mechanical Turk's secret was laid bare, it was incontestable that the Mechanical Turk did not have the sufficient functioning to qualify as an AGI.

Modern AGI research is proceeding under the same faith that once we stumble on general intelligence, the processes that yield the discovered general intelligence will satisfy intuition. The result of this lack of definitive, sufficient properties for AGI, is a disjointed research field that is pursuing a scattergun approach towards the Grand Ambition.

It is a matter of foundational importance to the research field to nail down, as much as possible, a firm understanding on both the necessary and sufficient qualities that will qualify an artefact as an AGI.

2.2. Unpacking the necessary properties for AGI from the Grand Ambition.

The difficulty in reaching a consensus on the necessary properties that would qualify an artefact as an AGI lay in the fact that the Grand Ambition of AI research was not suitably agenda defining. The Grand Ambition is to develop artefacts that possess general intelligence that is functionally comparable to human general intelligence. While this objective may seem straightforward, it is riddled with the potential for disagreements about definitions, inconsistencies in interpretation, and implications that are counterintuitive to the AGI agenda.

This section will highlight the challenges in interpretation and the potential for disagreement in the key words that are used to define the Grand Ambition. It will attempt to extract from their resolution the implied minimally necessary properties for an AGI artefact.

2.2.1. Intelligence.

The initial challenge in building a consensus for the necessary properties that will qualify an artefact as an AGI begins with adopting a standard definition of intelligence to inform the research field. The word 'intelligence' is prominent in the Grand Ambition. A necessary property for an AGI artefact is that it must be intelligent in some form that satisfies the Grand Ambition. It is therefore prudent to establish an understanding of the type of intelligence that will qualify an artefact as an AGI.

Despite its frequent usage in the research field, 'intelligence' is a term that proves difficult to define for broad consensus. Surveys of industry professionals have revealed hundreds of definitions of intelligence (Monett & Lewis, 2017), (Legg & Hutter, 2007). These definitions centre on intelligence broadly and that breadth is reflected in the diversity of definitions fostered. It would be unwieldy to parse through all the definitions to extract a common thread.

Given that our interest is in intelligence for artefacts, it would be useful to focus on definitions for the types of intelligence that should be sought in artefacts. There is a methodology to this decision beyond an interest in narrowing down the task. Coelho Mollo suggests that any definition of intelligence adopted for AI should be responsive to existing scientific practice and knowledge and should not be radically revisionist (Mollo, 2024). The AI research field has made some attempts to define the type of intelligence that it seeks.

Some prominent examples of attempts to define the intelligence that should be sought in artefacts include -

- Stuart J. Russell's definition of intelligence as actions of agents that are functions that input tuples of percepts from the external environment and produce rational actions based on these percepts (Norvig & Russell, 2009) .
- James S. Albus states that intelligence is the ability of a system to act appropriately in an uncertain environment, where appropriate action is what increases the probability of success, and success is the achievement of behavioural subgoals that support the system's ultimate goal (Albus, 1991).

- David B. Fogel states that intelligence is any system that generates adaptive behaviour to meet goals in a range of environments (Fogel, 1995).
- Herbert A. Simon and Allen Newell state that intelligence is any real situation behaviour appropriate to the ends of the system and adaptive to the demands of the environment (Newell & Simon, 1976).

The focus on artefactual intelligence has narrowed the task, but there appears to be more work to be done to establish a consensus.

A logical starting point would be to establish a definition of intelligence that seeks consensus on the aspects of intelligence that would be relevant and useful for AGI research. Coelho Mollo suggests that an adequate definition of intelligence should have epistemic distinctiveness in that it plays a distinctive theoretical and explanatory role in the relevant science. To meet this goal, a concept of intelligence must structure scientific theory and experimentation in productive ways, create promising avenues for research, and provide valuable tools for explaining, measuring, and modelling the phenomena of interest (Mollo, 2024).

AGI research not only aims at discovery, but also has practical aspirations. It aims to create a useful product. In aiming to replicate intelligence, AGI research aims to develop a form of intelligence that can be used. What is interesting about intelligence for AGI research is its practical applications. A careful examination of the various responses in surveys pertaining to the definition of artefactual intelligence reveals that intelligence is universally understood as a capability or set of capabilities that enable the recognition, engagement, and systematic resolution of new or existing problems. Take the examples we cited in the previous paragraph.

- Russell's definition of intelligence is actions of agents that are functions that input tuples of percepts from the external environment *(problem defining inputs)* and produce rational actions based on these percepts *(solutions*) (Norvig & Russell, 2009).
- Albus states that intelligence is the ability of a system to act appropriately in an uncertain environment *(problems),* where appropriate action is what increases the probability of success, and success is the achievement of behavioural subgoals that support the system's ultimate goal *(solution)* (Albus, 1991).
- Fogel states that intelligence is any system that generates adaptive behaviour *(solutions)* to meet goals in a range of environments *(problems)* (Fogel, 1995).
- Simon and Newell state that intelligence is any real situation behaviour appropriate to the ends of the system *(solution)* and adaptive to the demands of the environment *(problems)* (Newell & Simon, 1976).

As a further example, consider the Mechanical Turk. What was so impressive about the Mechanical Turk was its output. It could play chess, perform the Knights Tour, converse with its opponents, and identify attempts at cheating. All these exhibited behavioural outputs could be framed as appropriate solutions to problems that it was confronted with in its environment. Playing chess can be understood as solving the problem of winning in the specific rule-based game of chess. Performing the Knight's Tour is solving the problem of moving a knight across all the squares of a chessboard. Responding in a conversation could be defined as solving a problem of responding suitably to interlocution. The Mechanical Turk's capacity to attend to all these problems with appropriate output is what gave the Mechanical Turk the appearance of intelligence. Intelligent output can be described as an appropriate response to a problem.

A definition of intelligence for AGI research can thus be built around what is desirable about intelligence for the AGI researcher. The interesting aspect of intelligence for AGI research is the propensity to recognize, engage and appropriately work towards solving general problems in an environment. This adopted definition of intelligence as geared towards problem solving is in keeping with the thoughts of early pioneers in the AI research field who were labouring to realize a general problem solving program (Newell, et al., 1959)

2.2.2. Generality.

The name of the research field, Artificial *General* Intelligence, features the word 'general' prominently. As does the Grand Ambition of the research field. The Grand Ambition is to realize artificial intelligence that is functionally comparable to human *general* intelligence. It is evident that generality is a necessary component for success in the research field.

There is a well-defined understanding of generality in relation to cognitive abilities. Generality is the existence of a broad mental capacity that influences performance on other kinds of cognitive tasks. Charles Spearman who first suggested a theory for general intelligence argued there was a general factor 'g' which represented a common ability that was influential in tasks across different domains (Spearman, 1904). Godfrey Thomson proposed that rather than conceptualize 'g' as one thing, it would be better to think about it as collection of diverse skills needed to complete most intellectual tasks (Thomson, 1916). What is common to both definition is that generality is a skill or skills that are transferable across domains.

In stating that the goal is to realize general intelligence that is comparable to human intelligence, the Grand Ambition suggests that the generality that should be sought for AGI should be inspired by the notion of generality in cognition.

When viewed at face value, this demand for artefactual generality that is like cognitive generality does seem a perfectly straightforward. General in this case seems to refer in a very loose sense to an intelligence that is applicably broad and not bound. The adoption of the term into the AI research field is also motivated in part by the desire to contrast the field of research against artificial narrow intelligence and its boundedness to narrow applications (McCarthy, 1987). However, both ‘broad’ and ‘unbound’ are obscure descriptors. In what sense can intelligence be broad or unbound?

The most widespread understanding of the word ‘generality’ in AI is that it marks out problem-solving over multiple domains (Fogel, 1995), (Legg & Hutter., 2007), (Yang, et al., 2020). The interpretation of broad in this reading refers to covering a considerable number or a wide scope of things. There has been research in the field that has suggested useful ways to measure competency across the breadth implied by generality. This research has mainly focussed on identifying metrics that factor in the difficulty of the tasks across the domains (Hernández-Orallo, et al., 2021), (Morris, et al., 2024).

While it is useful to have an acknowledgment of what type of generality is sought and how to measure it when it is at hand, the definition of generality as competence across a breadth of domains is not suitably agenda defining. The definition appears to capture the spirit of what the AGI researcher means when they speak of generality, but it raises several questions about its practical application.

For example, if general intelligence is understood as problem solving competency over a considerable number or wide scope of problems, a literal reading of this demand would set a misleadingly low bar. An artefact with multiple domain problem-solving competency can easily be achieved by equipping an artefact with a wide repertoire of pre-defined problem-solving programmes. My computer is a Von Neumann machine that can execute as many tasks as are informed by the pre-installed programmes with which it is equipped. My computer has as broad a domain mastery as the programmes it is equipped with, and it has the potential to acquire as much more domain competency as the additional programmes I can fit onto it. Yet it is decidedly not an artificial general intelligence. There must be further demands on the type of generality sought by AGI beyond multiple domain problem solving competency.

There are further questions that an engineer who seeks to identify a target to work towards to can raise about defining generality as problem solving competency over multiple domains. What number of mastered domains qualify as general intelligence? At human levels is self-evidently anthropocentric, and potentially too high, while anywhere else seems arbitrary. How are we to define domains? Can subgoals be domains that account for generality even within a very restrictive main goal? For example, if an AI robot is built to make coffee, does its ability to grind the coffee beans qualify as a different domain from the ability to brew the coffee from the ground beans? All these questions highlight the challenges that must be resolved if generality is defined as multiple domain mastery and set as a necessary indicator of fulfilment in the research field.

One interpretation of generality that avoids all these concerns is defining generality as the ability to self-acquire problem-solving competency in new domains (Shanahan, 2015, pp. 5-6) (Hernández-Orallo, et al., 2021). Shanahan has suggested that we should think of AGI as an intelligence that is non-specialized, but can learn like a human to perform several tasks (Shanahan, 2015), (Morris, et al., 2024). In this definition the word 'general' refers to unbound in the sense that the entity is not limited or bound to one thing. It can learn to do new things. Contrast this with artificial narrow intelligence. ANIs are notoriously bound to one specific problem. They are rigid, and they are often brittle in the sense that they breakdown even with the slightest shift in the context of problem setting. General intelligence is unbound in the sense that it is not limited to solving one problem. It has potential to acquire problem-solving competency when faced with new challenges, or differently contextualized challenges.

It should be noted though, that under this reading of the word 'general' to mean something that is not limited or bound, the ability to learn is not unlimited. Any system's problem-solving capacity and capability are always going to be limited by design and resources. Because of the limitations occasioned by design and resources, intelligence or problem-solving efficacy is often specialized for native environments. A native environment is an environment in which the subject with intelligence has been optimized, either through adaptation or configuration to inhabit or to address problems. Efficiency in utilizing resources and the advantages of design in solving general problems in native environment appears to be a hallmark of general intelligence (Hernández-Orallo, et al., 2021).

If it is accepted that self-acquisition and refinement of new problem-solving competency is the most desirable part of generality, the questions raised earlier about generality as simply multiple domain mastery do not grate at the intuition anymore. The shortcoming of my Von Neumann machine is that by itself it cannot do anything. Its capability in problem solving comes from pre-built programmes, not self-acquired or learnt problem-solving efficacy. Someone else has solved the problem, encoded it, installed it, and the machine simply executes the solution. It does not have general intelligence, because even though it can execute several problem-solving tasks, it has not self-acquired the ability. As for questions about how many domains must be mastered for an artefact to be deemed an AGI, and what qualifies as a problem domain, they do not hold any value when the interest in generality is limited to an ability to self-acquire competency at meeting problems.

A suitable agenda defining definition of the desirable form of generality in AGI therefore should go beyond a mere requirement for cross domain problem solving. It should also specify an ability for the self-acquisition and refinement of problem-solving competency.

It is worth commenting here on certain game-playing agents that have shown competency in learning to play new virtual games despite not having been provided with the rules of the games. A prominent example is the MuZero algorithm that achieves this broad competency by combining tree-based search with a learned model, that enables it to play different complex, visually challenging games without needing explicit knowledge of their underlying rules (Schrittwieser, et al., 2020). MuZero managed to play 57 Atari games and equalled AlphaZero's accomplishments in Go, chess, and shogi (Schrittwieser, et al., 2020). These game playing algorithms appear to satisfy the adopted description of generality as the ability for the self-acquisition and refinement of problem-solving competency over both familiar and novel problems.

A case can be made that there is a distinction that can and should be drawn between the generality that is sought in AGI and the generality that is exhibited by these game playing agents. AI agents like MuZero exhibit generality in the sense that their training offers a solution with broad applicability across various situations in a narrow domain. The generality sought for AGI seeks knowledge that is adaptable, flexible and transferable from specific examples to new cases. Increasingly this distinction is being recognized as a distinction between generalization and generality (Hernández-Orallo et al., 2021).

Further explication might tease out this distinction more clearly. Agents like MuZero, that play multiple games are trained to acquire a model that is applicable across various simple virtual game scenarios. These agents process entire frames of pixels, and by interacting with possible changes in these frames in response to the agent's potential decisions, it learns the best policy, the likely outcome value, and the immediate reward. While processing entire frames of pixels appears daunting, in 2D games like Atari and the virtual environments for games like chess and go, it is relatively simple for AI agents. Indeed, agents like MuZero further simplify the environment by turning their observation into a 'hidden state' that is an even more simplified version of the current state (Schrittwieser, et al., 2020). These agents have generalization within their trained scope that allows them to generate an action-selection policy well enough over different relatively simplified games in a pixelated virtual environment that returns rewards under game rules. There is some generalization here, but it is in a very narrow domain. The same agent could not transfer these learnings to a slightly more complex environment. In fact to account for complex games like Dota 2 or StarCraft, the built agents had to be specialized and use the game's API to get direct information that bypasses the need for the agent's observation (The AlphaStar team, 2019), (OpenAI, 2019).

Game playing agents like MuZero have been equipped through training to have competency that is applicable across games with different rules only when they are presented in a simplified and narrow context (2D, pixelated, virtual environment). They have generality within a strict domain, but fall short of the broader, adaptable general intelligence envisioned by the Grand Ambition. The Grand Ambition seeks general intelligence comparable to human (or cognitive) intelligence. The flexible and adaptable general intelligence envisioned in the Grand Ambition should be transferable across domains and context.

A good north star definition for 'general intelligence' that conforms to the Grand Ambition is that general intelligence is problem-solving competency that is self-acquired, adaptable and refinable to address both familiar and unfamiliar problems across domains in a native environment.

2.2.3. Capacity.

The Grand Ambition of AGI research also makes explicit reference to comparability with human intelligence. A necessary property of the type of general problem-solving competency sought is that it must be suitably comparable to human problem-solving competency in some ill-defined way.

One interpretation of the Grand Ambition's demand for comparability with human intelligence is that a necessary quality for artefactual intelligence is that it compares favourably to some factor of human intelligence's capacity for problem solving. Capacity in this case is the computing power and/or processing resources that a system can leverage towards problem solving. The temptation would be to benchmark a necessary indicator of success in AGI research on realizing some pre-defined computational or processing capacity that stands in suitable relation to the capacity that supports human general intelligence.

The folly in benchmarking the success of AGI research on some necessarily realized computational or processing capacity is twofold.

First, there is no convincing evidence to suggest that increased computational power and processing resources are a guarantor for general intelligence. Those that subscribe to the idea that an AGI can be realized through brute force of computing capacity are most likely motivated by the fact that until recently, it has appeared to be the case that the more computational capacity was accessible to an AI model, the better it performed at complex tasks. Advocates for this line of argument might point at the improvements in the efficacy of large foundational models like ChatGPT against benchmark tasks that have been designed to indicate human performance. These models have become better at benchmark tests as their dedicated computing resources have grown. Taking Open AI's GPT models as an example, Epoch AI estimate that GPT4 used anywhere between 3,000 to 10,000 times more compute than GPT2 (Epoch AI, 2024). The increased compute registred noticeable improvements in benchmark performances. Benchmark performances equated GPT2 to a pre-schooler. The newer GPT4 model has performed comparably to a higher schooler on popular benchmark tests (Aschenbrenner, 2024). The improvements in performance have motivated and justified investments in compute for next generation models. As an example, Microsoft and Open AI are in the initial phases of constructing a $100 billion computing cluster to be brought online by 2028 for their next generation model (Reuters, 2024). By comparison GPT4 was trained on a $500m cluster. Leopold Aschenbrenner has made a case that we are on the path to building $1 Trillion computing clusters by the end of the decade (Aschenbrenner, 2024). Arguments have been made that if these investments in compute realize the same leap in performance from GPT2 to GPT4, the next generation models will perform comparably to PhDs at benchmark tasks and might very well be AGIs (Aschenbrenner, 2024), (Bubeck, et al., 2023).

However, one must be careful not to conflate efficacy over designed benchmark tasks with the aspirations of the Grand Ambition. Increasing computational capacity has not brought us closer to unlocking cross domain problem solving yet. Returning to the example of the large foundational models, despite their significant improvements in performance these algorithms are still fundamentally solving one problem. They are trained on large volumes of tokenized data, and they learn patterns from the training data that allows them to predict the next token. They are large and expensive artificial narrow intelligence models. A computationally potent ANI might be desirable in its own right, but it is not the goal of AGI research.

Too boot, recent developments indicate that the design of AI models might have more bearing on efficacy over complex tasks than brute force of computing capacity alone. The computationally frugal Liquid AI models, an artificial neural net which is modelled after the neural system of a nematode worm (Caenorhabditis elegans), performs comparably or better than models that were two-orders of magnitude its size at set tasks (Hasani, et al., 2020). The Liquid AI models are also capable of adjusting their parameters for success in response to changes in their environments. The Liquid AI architecture suggests that advancing problem-solving competency, whether that be in terms of addressing complex tasks or unlocking cross domain problem solving, might be addressed by the right design, rather than simply attaining a critical amount of computing capacity.

The second problem of benchmarking success in AGI on capacity is that it is not clear where we should set the necessary mark. The most obvious approach down this road would be to define a requirement for capacity against some yet unachieved computational capacity within AGI research. A benchmark within contemporarily accessible computing capacity is not sensible. It will raise the question as to why we have not achieved an artificial general intelligence artefact today. Any answer to this question must include a tacit acknowledgement that computational capacity alone is not a useful indicator for meeting AGI research goals.

There have been several arguments made that the benchmark for a realized AGI should be a computational capacity that matches human processing capacity. Many of the early predictions for the eventual realization of AGI are based on the assumption that it will occur when AI systems can match the computational resources of human mental processing (Vinge, 1993), (Chalmers, 2010), (Moravec, 1990), (Moravec, 1998), (Kurzweil, 2005). The argument made by advocates of this approach is that the point of realizing this benchmark is imminent. The evidence being the fact that leverageable computing capacity has grown at a predictable and steady rate as the cost of microchips drop and their efficiency grows. Moore's Law predicts future growth of computing power in microchips. The law suggests that computing power will double every two years. This prediction is based on the observation that the number of transistors that can be fitted on dense integrated circuit historically doubles every 18 months. Using Moore's law as predictor for growth in computing capacity, it has been estimated that we would be able to produce artefacts with human computing capacity in the near future (Moravec, 1998), (Kurzweil, 2005). People who cite this as a milestone often take it for granted that an artificial general intelligence will emerge or follow closely from the realization of this milestone.

The problem with basing a necessary indicator for AGI based on Moore's Law is that it is not a law of physics, and therefore it is not a dependable factor on which to pin success. Moore's Law is a projection of future growth based on historical growth. What is more, the projection is used to guide long term planning and set Research & Development goals within the semi-conductor industry. By serving as an industry standard aspiration, it has appeared stable because it is a self-fulfilling prophecy. That said, Moore's law is likely to eventually become undone. Eventually it will become harder and harder to make smaller and smaller microchips without a paradigmatic change in technology that accounts for the laws of quantum physics. There is no reason to believe that in the case of new microchip technology, if it ever emerges, that growth in accessible computing power will continue along the same trajectory that Moore's law defines. In fact, some powerful arguments have been made that we have already reached the limit of the law (Peper, 2017). The danger of setting a necessary indicator for AGI based on some yet unachieved amount of computing resources is that the target might remain out of reach. It is difficult to legislate for the future, so it is hard to say with any certainty how much computing capacity will be accessible to AGI research in the future. However, even if Moore's law or some other predictive model for growth in computing power was deemed reliable, it would not be sensible to lock out paths to general intelligence that might be computationally frugal.

Interpreting the Grand Ambitions demand for artefactual intelligence that compares favourably with human intelligence as a demand for comparability in computational capacity is misleading.

2.2.4. Capability.

Another often misleading interpretation for the Grand Ambition's demand for comparability with human intelligence is that it is a demand for problem solving ability that is suitably comparable in capability to human problem-solving. Capability refers to the skill, ability, or competency of a system in problem solving. Suitable comparability in this case will imply problem solving capabilities that match human problem solving in output.

This interpretation suffers from some of the same problems as the capacity interpretation. Chief among which is identifying where to place the necessary benchmark for capability that qualifies as AGI. The temptation is to read the demand for comparability with human intelligence as the indicative benchmark. However, it is not clear if this human-level benchmark for comparability is to be read as the upper limit or the lower limit.

If one were to consider human capability the upper limit against which comparability qualifies an artefact as an instance for AGI, it would be unnecessarily restrictive. It is broadly accepted that if AGI is at all viable, that it should and will eventually outperform humans in capability. Setting an upper limit on what one defines as AGI is potentially counterproductive. It would lock out high performing artefacts with super intelligence.

Adopting human capability as a minimum limit to qualify an artefact as an instance of AGI is just as flawed. Given that technology tends to start with limited capability and improves with the engineers' mastery of their relevant discipline, it is highly likely that if successful, the initial iteration of an artefact with general intelligence will fall short of the capability of human intelligence. It will only grow to outperform human intelligence with the engineers' mastery of their craft. Or through the AI's own autonomous self-development. Yet, in such a scenario it seems intuitive that the artefact will certainly be an AGI long before it has reached and surpassed human capability.

Closely related to the preceding point, one should also consider a school of thought in AGI research originated by Alan Turing which holds that to conform to the research brief, AGI should be realized with a capability that falls short of human intelligence. Human intelligence acquires its efficacy through experience and learning. If the intelligent artefact is truly functionally comparable to humans, it will acquire its efficacy in the same way, gradually through experience and learning (Turing, 1948 (1992)). A suitably designed intelligent artefact in Turing's child-AI mould will start with a capability that falls short of general human intelligence and grow in capability in a manner that is commensurate and limited by its design (Proudfoot, 2017). If proponents of Turing's child-AI are correct and it is the case that all AGI artefacts are realized as a *tabula rasa,* then intelligent artefacts at inception will be outside a human-level benchmark for success based on capability. It seems counterintuitive to deny that a Turing child-AI in its child state is already an AGI.

It should also be considered that most AGI researchers are not beholden to a strictly anthropocentric understanding of intelligence. Several researchers have adopted ambitions to seek artefacts with intelligence that have capabilities that could not be adequately compared to human intelligence. As an example, consider the different capabilities put towards animal navigation. Humans are good at visual spatial cognition, and it is the primary processes that informs unassisted human navigation. However, other animals use a host of other different cognitive skills that are informed by a range of diverse environmental cues. The input for these navigational solutions might include the location of celestial bodies like the sun, the moon and the stars, magnetic waves, olfactory stimuli, auditory stimuli, wind, light polarization etc. It is difficult to objectively compare these different forms of navigational capabilities adequately against human's capability (Shettleworth, 2010, pp. 261-4). Yet an artefact that could leverage any one of these other navigational capabilities would be just as impressive as one that had human visual-spatial navigational capabilities, and just as deserving of recognition as an intelligent artefact provided it could apply these capabilities adequately to solving problems in its native environment (Hernández-Orallo, 2017).

Adopting a strict comparability with human-level problem-solving capability as a necessary indicator for AGI is not ideal. Direct comparability to human-level capabilities is not necessary to qualify an artefact as an AGI. This is evident if one considers an early model AGI that has minimal functionality, or a Turing child-AI that starts of as a blank slate of capabilities but with the ability to learn when confronted with unfamiliar problems. In both these cases, these models would qualify as AGI artefacts long before they match human-level problem-solving capabilities. Adopting an interpretation of the Grand Ambition's demand for comparability with human intelligence as a demand for comparability in capability will also risk locking out AGI research projects that seek to realize capabilities that are not directly comparable to human capabilities, but none the less can be leveraged for effective problem solving.

This section has identified that AGI seeks an artefact with problem solving capacity that is general in the sense that it can self-acquire solutions to familiar, new, or differently contextualized problems in a native environment. The artefact should be functionally comparable to human intelligence in a way that is not strictly defined by its capacity or capability.

2.2.5. Functionality.

The Grand Ambition is to recreate intelligence that is functionally comparable to human general intelligence. The word 'functionally' in the brief appears to be doing important work. The implication of the word is that the goal for the AGI researcher is to recreate or at least mimic the functionality of human general intelligence. Problem-solving efficacy and generality are desirable only in as far as they are supported by functionality that is comparable to human problem solving. An AGI artefact ought to be judged for sufficiency on the functionality of the general intelligence it realizes in relation to human intelligence.

However, even with a reading of the Grand Ambition that centres functionality, there is room for disagreement. The dispute turns on two interpretations of the *ad litteram* definition of 'functional'. In purely conceptual terms, the word 'function' can be understood as a noun or a verb. Subtle differences in the definition of the two has had an outsized bearing on how the AGI agenda should be perceived and pursued.

i. Output-centric AGI.

The word 'function' understood as a noun means an action or end that is natural to the purpose of a thing. If the word 'functionality' in the Grand Ambition is derived from the noun function, then any realized artefact that outputs ends or actions that a human general intelligence would output if operating optimally with the same resources would be functionally comparable to human intelligence. The AGI researcher who reads a noun in the word 'function' is only interested in ensuring their engineered system generates what is deemed the correct output. The output-centric engineer interprets the Grand Ambition as a call to build artefacts that output solutions to general problems in their native environment that are comparable to human solutions. In this case, the assumption is that the hallmark of human intelligence is rationality. So, the output-centric engineer seeks to build an artefact that outputs optimally rational behaviour, where the rationality of the behaviour is judged from the perspective of a human observer.

The output-centric approach is the most common interpretation of the Grand Ambition in AGI research and development. Most AGI projects are attempting to build systems that consistently output what is deemed ideally rational behaviour as judged from a human perspective. The AI artefact itself is not expected to have its own subjective perspective on rationality, at least not one whose ends should be pursued or prioritized. Under this approach, the processes that realize the output are of trivial importance. What matters most is the correct output from input. The artefact is expected to output actions that conform to behaviour a human would deem rational given its input and resources.

ii. Process-centric AGI

The word ‘function’ understood as a verb means to work or operate in a specific or particular way. If the word ‘functionally’ in the Grand Ambition is derived from a verb, it implies that an artefact that is functionally comparable to human general intelligence would have to instantiate processes that are specific to human general intelligence. The AGI researcher who reads the word ‘functionally’ in the Grand Ambition as derived from a verb is interested in recreating in an artefact, the correct processes that yield general intelligence in humans.

The process-centric engineer understands the Grand Ambition as mandating an artefact that realizes an appropriate output to general problems in its native environment through functional processes that are comparable to human problem-solving states and processes. The goal under this interpretation is to identify the right processes that yield intelligent behaviour and to recreate or rebuild those processes in the artefact. The output-centric engineer is not seeking to prescribe problem-solutions to the artefact, or to impose their rationality on the artefact. Rather, they seek to realize a system that has the correct functional processes that will enable in it its own subjective states that will yield problem solving. The view is that if these processes are realized correctly in a manner that is comparable to human processes, the output that will follow will be comparable to human output.

The process-centric interpretation of the Grand Ambition has the firmest grip on the popular imagination. Depictions of intelligent artefacts in popular literature and the recent popular narrative on AGI seem to stem from this interpretation. The understanding is that AGI artefacts will be machines with minds, i.e., machines that have the processes to generate a subjective rational perspective that drives behaviour. If done correctly, the artefact is expected to have states equivalent to thoughts, beliefs, desires, moods, etc., just like the states in human minds that influence behaviour.

At face value, the Grand Ambition has left open the two possible approaches towards AGI. The first is the output-centric approach. It interprets the ambition of AGI as building any system that outputs problem-solving behaviour that a human would deem rational. The emphasis on this approach is on establishing the correct input to output stream that produces result that are comparable to human intelligence. The output-centric engineer is agnostic about system processes. The second approach is the process-centric approach. It interprets the ambition of AGI as realizing in an artefact with the right human-comparable processes that yield problem-solving behaviour. The process-centric engineer places the emphasis on the right processes. They believe that the right problem-solving behaviour will emerge from the right processes.

So which approach is truest to the brief of the word ‘functionally-like’ in the Grand Ambition? Both approaches yield a necessary feature of general intelligence i.e., the right output, which in this case is rational problem-solving behaviour.

As a philosopher, one is tempted to side with the process-centric meaning as the closest to the goals of artificial general intelligence. One good reason for this perspective is that while both approaches can be subsumed under the philosophical definition of functionalism, the process-centric view appears to be a better alignment.

In the philosophy of mind, functionalism is the view that what defines a mental state as a specific type is not its constitution, but rather the role it serves or how it functions within the cognitive system (Levin, 2023).

The output-centric engineer believes that the task at hand is realizing any system that yields an output that is comparable to human outputs that would be deemed general intelligence. The output-centrist recognizes that they are engaged in a functionalist pursuit. Their version of functionalism is non-committal about what type of processes are necessary to realize the correct output. They believe that any method that yields the right human comparable behaviour serves an appropriate functional role.

Conversely, the process-centric engineer also values the importance of producing outputs comparable to human intelligence, but they place primary importance on discovering the correct process that produces these outputs. This process, in their view, should mirror the process underlying intelligent behaviour in humans. For the process-centric engineer, AGI should be functionally comparable to humans, not only in output but in how outputs are produced. They believe there is a correct process for intelligence, even if the process can be realized with different underlying materials.

The process-centric view aligns more closely with philosophical functionalism. Functionalism allows for multiple realizability but underlines the importance of the processes that are enabled by the state. A commonly used example is one of pain experienced from a burn. A functionalist theory might define this pain as the state triggered by bodily injury from a heat source that indicates the injury’s location and informs actions to withdraw the affected area from the heat source. Here, functionalism allows flexibility in constituent materials (for example it might be C-fiber stimulation or another form of experiential indicator) but emphasizes the functional role of pain. The theory’s focus is on the processes that pain manifests rather than the specific materials used. The process-centrist’s approaches the problem by understanding the process through which pain realizes the correct output and replicating the process without a commitment to constituent materials. The output-centrist in being agnostic about discovering processes and focusing instead on replicating input-output mapping is also committed to material flexibility. However, this approach is not as good a fit for philosophical functionalism, because the process that sandwiches the input and output is just as important in defining a function. For example, not any process that maps contact with heat and actions to retreat from heat qualifies as pain from a burn. It is arguable that to qualify as pain, the state must enable the output in the right way. The state of pain should be in a causal relation with the burn, it should indicate the location of the burn, it should indicate an unpleasant state, and it should motivate the action.

Given that the output-centric approach is the most prevalent in practical AGI research, it seems that the philosophical definition of functionalism has not had strong enough influence on the engineering side of the research field. A thought experiment can illustrate why the output-centric view is flawed and does not sufficiently reflect the ambition of replicating intelligence that is functionally comparable to human-like intelligence.

2.3. Tracing the sufficient conditions for general intelligence.

The output-centrist are either overtly or tacitly committed to the notion that intelligence that is functionally comparable to human intelligence is sufficiently judged in a computational model by its output. Under this reading, AGI research seeks to realize artefacts that output actions or behaviour that are deemed rational, where rationality is judged from a human perspective. However, questions can be raised about whether output that is deemed rational is sufficient for defining intelligence.

A thought experiment can help highlight that there are more demands to be made in defining intelligence than the mere output of outwardly rational actions (behaviour). The thought experiment will help to surface what conditions would need to be met to sufficiently qualify seemingly rational output as intelligent.

Consider the following thought experiment.

Thought Experiment: Algebra Examination.

A school sets a short multiple-choice examination in elementary algebra. Examinations are meant to gauge a student's proficiency in a subject by assessing their problem-solving competence in the subject. In this algebra examination, the students are required to find values for abstractions represented by letters in equations by computing their value using only their knowledge of the mathematical symbols and the rules they engender. To pass the examination, the students must get at least 80% of the answers right. The rationale is that an output of 80% will demonstrate a satisfactory problem-solving competence in elementary algebra.

A student named Johnny has the goal to pass the examination. However, he is completely unprepared for the examination. He has missed all the lessons relating to algebra and as a result he does not have even the most basic know-how of solving an algebraic equation. Nevertheless, he attempts the examination. His strategy is to attempt a correct guess from the multiple-choice answers provided. It works. Without reading any of the questions, Johnny manages to successfully guess enough of the answers to get a high passing grade.

In this case, Johnny the student has produced output that would be deemed rational by a human observer. By scoring 80%, his answers would be deemed a suitable demonstration of problem-solving competence in algebra by the examiners. Johnny has satisfied a necessary condition for demonstrating problem-solving competence in algebra. He has produced the right output. However, Johnny has NOT satisfied sufficient conditions for demonstrating problem-solving competence in algebra. Johnny could not possibly be said to have output sufficiently competent problem solving. The indication here is that there are further demands that can be made of the correct output for it to be sufficient as an intelligent output. By itself, the right output is not a sufficient indicator for intelligence.

By understanding what irks us with Johnny's approach to the examination we can identify at least three properties that jointly need to accompany the right output as sufficient conditions for problem solving competence.

2.3.1. Sufficient condition 1 – Access to Relevant Input.

The first problem with Johnny's approach to the examination is that Johnny's answers, though appearing outwardly rational, are not a function of relevant input. Johnny did not bother to read the questions before he attempted to solve them. Johnny has not engaged appropriately with the problem he is trying to solve. Even if we are to constrain ourselves from commenting on his method for now, it can be agreed without controversy that for Johnny to meet the bare minimum for intelligent output, his answers should at the very least be some sort of function of the questions set.

Intelligence is directed towards recognizing problems in the world and engaging in problem-solving. Problems from the world are defined by relevant input. Input is information from the world that defines the state of the world. Before a problem can be solved, there must be the reception of appropriate input that defines the relevant problem in the world. The problem precedes the solutions.

If Johnny did not bother to read the questions before he attempted to solve them, then he could not be said to have engaged in appropriate problem solving for the algebra examination. Johnny somehow managed to produce the right output, but without the right input to define the problem. Though his output is correct, it is not sufficient for demonstrating intelligence. Any definition of intelligence as the outputting of correct actions will be incomplete without stating that the output should be a function of meaningful input. The relevant input in this case is information that defines the problem that must be solved.

2.3.2. Sufficient condition 2 - Appropriate function.

The second problem is that Johnny has not applied an appropriate function to realize his output. Simply judging intelligence as the correct output from any functions run on relevant input falls short of a satisfactory definition of intelligence. To elucidate this point, one can make a small adjustment to the thought experiment.

Imagine that Johnny did take the time to read the set questions in the examination. In this scenario, Johnny decides to base his answers on clues he thinks are hidden within the questions. He chooses from the multiple-choice answers, A, B, C, D, or E, based on which of these letters that he sees first in the question. For example, for the questions below, Johnny chooses the answer A, because the letter 'A' appears first in the question.

A number X is divided by four, after which three is subtracted to leave a result of zero. What is X?

A. 12
B. 40
C. 31
D. 02
E. 05

Even though he has used a contrived function to generate each of his answers, he nevertheless manages through sheer fortune to get a passing grade over 80%.

In this case, Johnny has the correct input, some form of function processing the input, and the correct output. Yet it would still be inadequate to call his actions sufficient for intelligence. The problem here seems to be with the nature of the function applied to the input. It is apparent that just any function that results in a correct output is not sufficient for intelligence. A function deemed sufficient for intelligence must be appropriate for the task at hand. To be deemed intelligent, the system must output the correct action, based on an appropriate function, run on relevant input that define the problem.

What qualifies as an appropriate function? There are at least three demands to be made of a function to deem it appropriate as a problem-solving function.

I. It must be creditable with success for generating the right output.

The first demand is that an appropriate function must be creditable with success in generating the right output. Johnny read clues in the questions that he thought pointed him to the correct answers and he happened to pass the examination. Johnny's success in the examination was entirely reliant on luck. While Johnny engaged some cognitive function to generate his answers, his passing of the examination could not be credited to the cognitive function. He was successful by mere happenstance. An appropriate function must be one that can be credited with success for generating the correct output.

II. It must be reliable for solving other problems in the domain (domain mastery).

Secondly, an appropriate function must be reliable for solving similar problems in the same domain. Again, one can make a small adjustment to the thought experiment to tease this point out.

Suppose that the examination setter had purposefully hidden clues to the answers within the examination question, perhaps as a private joke. Johnny's method by pure happenstance had cracked the code.

In this case, the method would qualify as a function that could be credited with Johnny's success. However, it is still apparent that it is not intelligent behaviour. One's intuition in this case, is piqued because one can raise questions about the reliability of his approach. His method might have worked for this examination, but it would certainly not work if a less jocular setter set the examination. Once again it is hard not to discount the oversized influence of luck in accounting for Johnny's success. He just happened to decide on a method that coincided with the setter's in-joke. In this case his function might be suitable for solving a particular kind of clue-based tasks, but it is not the correct function for the domain of algebra. An appropriate function must be a reliable approach to solving similar problems in the same domain. It is safe to say that Johnny's approach is not a reliable way to tackle algebra problems.

III. It must be adjustable and adjusted to fit context.

The third demand for a function to qualify as an appropriate function in problem solving is that the function must be adjustable to fit context. Returning to the example, if one imagines that the extent of Johnny's algebraic application for the rest of his life will be limited to answering questions set by the same setter, who is chronically bedevilled by the same sense of humour, then in that case one would have to concede that Johnny's method is both reliable and could be credited with his success in algebraic pursuits. Yet, it is still difficult to think of Johnny's behaviour as intelligent.

The problem in this case is that Johnny's approach to algebraic problems is context bound. In the A.I. parlance his function would be called a brittle algorithm. A brittle algorithm is one that works very well in a very narrow context but falls apart with the slightest change in the context of its input. The way algebraic problems are presented in the world will vary. Consider that elementary algebra is essential for the study of mathematics, science, and engineering. It is also applied in medicine, economics, and computer science. Johnny's algebraic function could not possibly be useful in the new context of any of these fields. His algebraic method is limited to the very narrow context of algebra examinations set by a particular setter, provided that the setter retains the same sense of humour applied in the same way. A sufficiently intelligent response demands the capability to generate a useful solution, and to be able to adjust the solution to fit different contexts. For intelligence, an appropriate function must be adjustable to solve the same problem in different context.

Johnny's problem-solving function for the algebra examination yields the right output from the right input, but it is not sufficient to qualify as intelligence because it is not derived from an appropriate function. An appropriate function for intelligence must have three features, it must be creditable with success for the solution, reliable, and adjustable to fit context.

An intelligent system is expected to have the correct output, based on a function that is reliable, adjusted to fit context, and that can be credited with success for generating the output based on relevant input that defines the problem.

2.3.3. Sufficient condition 3 – Self-sufficiency.

One can make one further demand on the correct output for it to be deemed a sufficient indicator of intelligence. In addition to the right input and an appropriate function, the correct output must be generated with an appropriate degree of self-sufficiency. The demand for self-sufficiency as a critical component of intelligence can be illustrated by making another adjustment to the algebra examination example.

In this version of the thought experiment, Johnny has devised a new scheme to pass his algebra exam. Johnny has attached a recording device to his eyeglasses, and it relays information Johnny reads from the paper to Sandy, who is in a remote location. Sandy is good at algebra, and she manages to apply the appropriate function to the input and relay the answers back to Johnny through an earpiece. Johnny then writes down the answers he hears on the examination script. Johnny passes his exam.

In this case Johnny has the right input, an appropriate function, and the correct output, and yet it would be uncontroversial to state that Johnny has not exhibited adequately intelligent output.

The problem is that Johnny cannot be credited with success in passing the examination. Indeed, if his scheme is discovered he would be penalized for cheating, because someone else deserves credit for the work he is passing of as his own. A requirement for Johnny's output to be deemed an appropriate output is that Johnny should be creditable with the output. To do so, Johnny needs to exhibit an appropriate amount of agency in generating the appropriate function for his correct output. In the last paragraph it was stated that an appropriate function is one that is reliable, adjusted to fit context, and that can be credited with success for generating the output. An appropriate degree of agency is when a function that meets all these criteria can be rightfully credited as the product of an intelligent agent's self-owned processes.

The Algebra Examination thought experiment has demonstrated that seemingly rational output while being necessary for intelligence, by itself is not sufficient to define intelligence. For rational output to qualify as intelligent action, it must be accompanied by three jointly sufficient properties.

1. The output must be processed from accessed meaningful input i.e., percepts that define the problem that must be solved.
2. The output must be processed by way of appropriate problem-solving functions i.e., function that is reliable, adjusted to fit context, and that can be credited with success for generating the correct output.
3. The output processing must be credited to the agent i.e., the appropriate function must be instantiated and attributed to self-autonomy.

The problem with the output-centric interpretation of the Grand Ambition is that it centres a necessary quality of intelligence, but it falls short of defining sufficient qualities for intelligence.

The output-centric interprets the goal of AGI as building an artefact that outputs actions that are deemed rational when judged by a human observer. An AGI project labouring under the output-centric interpretation is likely to miss critical features and properties that are important for an artefact to qualify as generally intelligent.

The process-centric interpretation of the Grand Ambition is not at risk of omitting both the necessary and sufficient qualities of intelligence. The process-centric approach focuses on modelling the processes of human intelligence that yield the right output. The approach will meet the sufficient conditions for general intelligence because if successful, the artefact will function under the same principles that govern human intelligence. If done correctly, the artefact will have functional equivalence with human general intelligence.

2.4. Weighing the Grand Ambition's definition of intelligence against desiderata for general intelligence.

The adopted definition of general intelligence in the Grand Ambition mould is problem solving competencies that are general in the sense that the artefact can self-acquire, adaptable, flexible and transferable solutions to familiar, new, or differently contextualized problems in its native environment. These solutions are output in the form of actions or behaviours. The artefacts problem solving competencies should be functionally comparable to human (cognitive) intelligence in a way that is not strictly defined by its capacity or capability but can be sufficiently demonstrated through actions that are processed from accessed meaningful input, processed by way of appropriate functions, and with an autonomy that is creditable to the artefact itself.

This definition of general intelligence is in conformity with the desiderata for intelligence that are proposed by Dimitri Coelho Mollo (Mollo, 2024). In the quest for establishing a consensus for a definition of intelligence, Mollo argues that it is essential to identify a notion that can explain a broad range of intelligence across different systems, including both biological and artificial intelligence. Key attributes for a good definition of intelligence should include (Mollo, 2024):

- **Species-neutrality**: It should not favour one species (especially humans) over others and should be applicable to both biological and artificial systems.
- **Origin-neutrality**: It should not limit intelligence to biological systems, allowing for the recognition of non-biological forms of intelligence, such as artificial systems.
- **Multiple realisability**: Intelligence may arise from various underlying mechanisms, both biological and artificial, so the definition should allow for different forms of realization.
- **Epistemic distinctiveness**: The definition should add value to scientific theories, offering a distinct way of categorizing and explaining intelligent behaviour.
- **Responsiveness to scientific practice**: The definition should build on existing knowledge and practices without radically revising them.

One might be tempted to argue that the adopted definition for general intelligence in the Grand Ambition violates the first desideratum for species neutrality. However, it should be noted that the use of the words ‘human comparable’ in reference to AGI in this thesis is in the interest of keeping conformity with the language that has become standard in AGI research. The research has recognized that this is a case of blatant anthropomorphism and has highlighted in this chapter and the previous one, that what is interesting about the comparability to human intelligence is common to most cognitive intelligence. Indeed, this chapter stressed that human benchmarks in capacity or capability are not useful in defining intelligence in the Grand Ambition. The comparability that should be sought should be in functionality that includes accessing meaningful input, processing it by way of an appropriate function, and doing so in a manner that is creditable to the artefact itself. These abilities are not unique to humans. There is some congruence here with Hernández-Orallo who adopts a single definition for both the intelligent behaviour in biological and computer organisms as a set of interactive systems that take in inputs, produce outputs possibly asynchronously through bodies and sensors (Hernández-Orallo, 2017).

There is nothing else in the Grand Ambition's definition of general intelligence that violates Coelho Mollo's desiderata for intelligence.

In addition, to the desiderata for a good definition of intelligence, Coelho Mollo identifies features of intelligence that many scientists agree on and therefore should be included in defining intelligence. These include features like generality (displaying broad competence), flexibility (adapting to new situations), goal-directedness (pursuing appropriate objectives), and adaptivity (learning from past experiences). These core traits should guide any development of a unified and scientifically fruitful characterization of intelligence (Mollo, 2024).

Once again, the intelligence defined by the Grand Ambition satisfies Coelho Mollo's views on what should be included in a definition of intelligence. The Grand Ambition defines intelligence that features generality, broad domain competence and the flexibility to acquire new problem-solving capabilities. It was also argued that this generality should differ from the generalization of game playing algorithms that are trained to solve broad but domain restricted problems but lack the adaptivity to transfer these learnings to recontextualized or different domains.

In subsequent chapters we shall explicate how the demand for autonomy under the Grand Ambition is related to goal-directedness. This thesis is interested in a particular type of goal-directedness, which is the capacity for intentional actions (Pezzulo & Castelfranchi, 2009). Intentional actions are the process through which the mind exercises its autonomy. The Grand Ambition mandates intelligence that is like the mind, therefore it implies intentional agency.

2.5. The Mechanical Turk and the conditions for AGI.

This chapter has identified both the necessary and sufficient qualities that an AGI project should aspire to realize in an artefact in the mould of the Grand Ambition. Armed with this informative and instructive definition of an aspirational AGI artefact, it is worth examining what it is about the Mechanical Turk that made its more studious observers question whether it was truly intelligent. The necessary and the minimal sufficient properties for AGI spelled out in this chapter can be used as tools to check if the Mechanical Turk or any other artefact with aspirations for general intelligence qualifies as an AGI artefact.

The chapter identified that the necessary properties for an AGI artefact is problem solving competencies that are general in the sense that the artefact can self-acquire solutions to familiar, new, or differently contextualized problems in its native environment. These solutions are output in the form of actions or behaviours. The artefacts problem solving competencies should be functionally comparable to human intelligence in a way that is not strictly defined by its capacity or capability.

The Mechanical Turk ticked all these boxes. It could certainly solve problems. It could play chess, perform the Knights Tour, converse with interlocutors, and identify attempts at cheating. All these actions and behaviours can be described as solutions that are output in the form of behaviour in response to problems in its native environment. The Mechanical Turk also demonstrated some evidence of the generality that is sought in AGI. The ability to apply solutions to new or differently contextualized problems. This is evidenced by the fact that it could respond appropriately to new conversational topics and to different attempts at cheating. Its responses to problems could certainly be deemed comparable to human intelligence in a manner that is not strictly defined by its capacity or capability. Take for example the way it conducted its conversations. It communicated through a letter board. This is certainly not directly comparable in capacity or capability to human vocal speech, yet in being able to respond appropriately to conversation, it is comparable in output.

In addition to the necessary qualities for AGI, the chapter identified properties that are sufficient for an artefact's output to be deemed intelligent. The actions must be processed from accessed meaningful input, they must be processed by way of an appropriate function, and the appropriate function from accessing input to generating output must be creditable to the artefact itself.

Together with the right output as a necessary future of intelligence, these three jointly sufficient features comprise the necessary and minimally sufficient properties for the Grand Ambition. They are minimally sufficient in the sense that they are the absolute minimum properties to qualify an artefact as intelligent in a manner that is functionally equivalent to human general intelligence. One could certainly imagine an intelligent artefact that can do more than the minimally sufficient properties, but to do less is to fall short of the requirements for the general intelligence envisioned by the Grand Ambition.

There is an intuitive understanding of the necessary and sufficient conditions for intelligence. Take for example the definitions of intelligence that were highlighted earlier in the paper. All of them hint at the identified sufficient qualities in their descriptions.

- Stuart J. Russell understands AI intelligence as actions of agents *[autonomy]* that are functions *[appropriate functions]* that input tuples of percepts from the external environment [input] and produce rational actions based on these percepts *[output]* (Norvig & Russell, 2009).
- James S. Albus states that intelligence is the ability of a system *[autonomy]* to act appropriately in an uncertain environment *[input]*, where appropriate action is what increases the probability of success *[appropriate function]*, and success is the

achievement of behavioural subgoals *[output]* that support the system's ultimate goal (Albus, 1991).

- David B. Fogel states that intelligence is any system *[autonomy]* that generates adaptive *[appropriate function]* behaviour to meet goals *[output]* in a range of environments *[input]* (Fogel, 1995).
- Simon & Newell state that intelligence is any real situation behaviour *[output]* appropriate to the ends of the system *[autonomy]* and adaptive *[appropriate function]* to the demands of the environment [input] (Newell & Simon, 1976).

To be considered truly intelligent, the Mechanical Turk must have satisfied both the necessary and the three minimally sufficient conditions of intelligence. It has been demonstrated that the Mechanical Turk did well with the necessary demands for intelligence. It generated the right output to general problems in its environment. But the Mechanical Turk fell short of meeting at least one of the three jointly sufficient conditions for intelligence. Since the three conditions are jointly sufficient for intelligence, The Mechanical Turk did not meet minimal sufficiency to be deemed intelligent.

The Mechanical Turk met the first of the jointly sufficient demands for intelligence. It generated its output from the correct accessed input. This is evidenced for example by the fact that it could play chess. Chess is a game that needs a player to respond appropriately to opponents moves. The fact that the Mechanical Turk could play chess implies that it surveyed the state of play and responded appropriately. It is unlikely that the Mechanical Turk was generating random actions that just happened to fit seamlessly with the state of play.

The Mechanical Turk also satisfied the second jointly sufficient condition for intelligence. The Mechanical Turk was applying an appropriate problem-solving function to generate its output. Its output in the various tasks that it performed appeared to be generated reliably and in fit with context. If the observers at the time took an interest in investigating the guts of the Mechanical Turk to determine whether the solutions the contraption generated were creditable to its functions, they would have been confronted with diversionary cogs and wheels that were meant to suggest that they were responsible for the output. It is now known that the Mechanical Turk's secret was a chess-master hidden within its guts. Even with this secret exposed, the Mechanical Turk still satisfied the second sufficient demand for intelligence. The hidden chess-master would have approached problem solving in chess, the Knights-Tour, and conversing with appropriate functions. The Mechanical Turk's actions were therefore generated by appropriate functions.

However, it is in meeting the third jointly sufficient condition for intelligence that the Mechanical Turk fell short. The third jointly sufficient demand is that the problem-solving functions should be creditable to the system itself. An extrinsic agent controlled the Mechanical Turk, even if it was the case that the agent was located within the guts of the system. The motivation to locate agency that qualifies intelligence was certainly the instinct that propelled Robert Willis to inquire from whence the contraption's 'promethean heat' could be found. Willis was labouring under the intuition that if credit for the actions of the system were to be located anywhere other than the system itself, it would preclude the system from qualifying as intelligent.

The Mechanical Turk falls short of this demand because its actions are made and are taken as a rational response from the perspective of an extrinsic agent concealed within the Mechanical Turk. This remains true even if we imagine that in a later model of the Mechanical Turk, the extrinsic agent, could control the contraption remotely instead of being hidden in the contraption. Or even in the case that the extrinsic agent had such foresight that he could predict every scenario the Mechanical Turk would be in and programmed responses to how the contraption should proceed in each case. In all these scenarios, the Mechanical Turk is acting under extrinsic agency. It is incapable of actions that can be deemed suitably creditable as self-sufficient.

The next chapter will identify why many computational approaches to AGI will run into the same problems that afflict the Mechanical Turk in meeting the third jointly sufficient requirement for AGI.

Chapter 3: On computational models of AGI, and a case for the replication of cognitive processes in AGI.

"We're like the man who climbed on a chair and declared he was a little closer to the moon."

What Computers Can't Do (Dreyfus, 1978)

3.0. The QTM, The AIXI, LLMs and the Grand Ambition.

The main thrust of AGI R&D, an agenda with the grand ambition to build machines with intelligence that is functionally comparable to human intelligence, has been pursued as a branch of computer science called Artificial Intelligence (AI). AI research in computer science was inspired in great part by Alan Turing's work in describing what came to be known as a Universal Turing Machine (Turing, 1936). A Turing machine is an abstract mechanism that computes by manipulating input symbols on a strip according to a set of pre-formulated rules to produce a desired output.

AI research has inspired computational theories of mind, which suggest that the human mind shares essential features with a Universal Turing Machine (UTM). A UTM is a programmable, general-purpose computer capable of simulating any other Turing machine. Supporters of this idea are known as computational cognitivists. The success of AI research at matching human intelligence at narrow tasks has led to optimism that computation can also yield success in solving for general intelligence.

On the surface, optimism that computational theory explains the mind and is therefore a path to artificial general intelligence appears to be warranted. There are unmistakable parallels between computation and cognition. Computation is the process through which a system uses algorithmic processes to simulate, describe, or project the transformation of symbolic information that represents states of affair or knowledge from the world. Similarly, the mind receives information from the world, represents it in some mental symbolic form, processes it through internal movements and manipulations to produce an output that represents or responds to transformations, projections, or knowledge of the world. The mind, the quintessential model of general intelligence, appears to be doing computation. If computation can yield general intelligence in a mind, it should therefore be the right path to artefactual general intelligence.

However, there are also unignorable differences between the mind and computation. The most glaring difference being that the mind hosts qualitative states, a phenomenon that does not appear to have any direct parallels in computing. Critics of computational theory of mind have argued that the processing of the mind, invigorated as it is by phenomenal experiences, is incomparable to the enervated syntactic manipulations of computations. They hold that computation is not the same and could never replicate the phenomenally tinged processes of the mind. Their position is that the assumption that artificial general intelligence is imminent based on ground covered by purely computational processes is a first-step fallacy. Hubert L. Dreyfus compared the optimism that computational models will bridge the gap to cognitive general intelligence based on the success achieved in narrow intelligence, to a man who supposes that he is close to the moon because he stood on a chair (Dreyfus, 1978, p. 12).

Computational cognitivists have remained bullish that the mind is doing enough to support an analogy with computation. They maintain that the right type of syntactic processing can yield general intelligence. They might highlight the Church-Turing-Deutsch principle (CTD) as proof that computational approaches can realize an AGI artefact. David Deutsch made the case for the physical principle that “every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means (Deutsch, 1985, p. 97). Those who see the Church-Turing-Deutsch principle (CTD) as proof that a wholly computational approach can realize an AGI artefact are making the argument that; Cognitive general intelligence is realized in a physical system. The CTD principle states all physical systems can be simulated through computational processes. Therefore, computational processes can realize an artefact with general intelligence.

Computational cognitivists currently lack a physical artefact to support their belief that computation can lead to AGI. However, there have been theoretical models proposed as potential candidates for AGI. Two such models, that can be said to embody principles that can realize general intelligence, are a Quantum Turing Machine built to simulate a cognitive general intelligence system and Hutter’s AIXI machine. More recently, some have suggested that with a few improvements in computational power and algorithmic efficiency, commercial large language models (LLMs) could also represent a form of AGI (Aschenbrenner, 2024), (Bubeck, et al., 2023).

David Deutsch suggested The Quantum Turing machine (QTM) as an example of a universal Turing machine (Deutsch, 1985). In formulating the proof for the CTD principle, Deutsch recognizes that the natural world is continuous while a computer like a Turing machine computes discrete items. The discrete nature of computational communications has certain limitations in richness and speed in computing continuous properties. So, he described an abstract Universal Turing Machine that uses quantum computing to enable a more comprehensive capture of the intricacies of the world and to enhance computational capabilities. Apart from its reliance on quantum processes to represent the complexity of natural systems and inject dynamism to their processing, the QTM can be thought off as operating under the same principles as a classic Turing machine. The machine is fed symbols, manipulates the symbols according to user specified instructions, and generates an output. Such a simple system could simulate any finitely realizable physical input-output system, within the bounds of practicability and constructability. It should follow therefore that a QTM could be built to simulate a cognitive general intelligence. It would be a QTM-AGI.

Marcus Hutter's AIXI system is another example of a general computing machine. The AIXI has been proposed as a theoretical formal system for artificial general intelligence (Hutter, 2000). The AIXI is a reinforcement learning agent that is built to maximize rewards from an environment. The machine considers every computable hypothesis in the environment, it then scans its several installed programs and evaluates how much reward each program would return. It weighs the programs according to its fit to the actual environment and selects a program that has the best ratio of economy in terms of computational resources to reward. Hutter has used his theoretical AIXI system to show that it would be remarkably successful at optimizing input – output solutions in several scenarios (Hutter, 2007).

Commercial Large Language Models (LLMs), such as ChatGPT and Google's Gemini, are natural language processors that use advanced statistical models to perform tasks like content generation, answering questions, translation, summarization, and text completion. These models are trained on massive datasets of natural language which enables them to recognize patterns and connections within natural language. The data is tokenized into smaller units, allowing the model to identify patterns in the data at a granular level. LLMs employ a type of neural network architecture called a transformer model to process and learn from the data. After the initial training, they are often fine-tuned on specialized datasets to enhance their performance on specific tasks or areas of knowledge. While LLMs can generate human-like responses based on the text they have processed, it is important to recognize that they do not possess the ability to think or feel like humans. However, it has been suggested that through brute force of computational power and some algorithmic tweaking to improve efficiency that LLMs could become AGIs (Aschenbrenner, 2024), (Bubeck, et al., 2023).

To determine whether computation is indeed a viable path to AGI, one could begin by examining whether these aspirational contraptions are truly AGIs. Do they satisfy the Grand Ambition of AGI research? This chapter will argue that they and other models like them that pursue an output-centric approach to AGI are conceptually flawed. They fall short of general intelligence because they are incomplete models of general intelligence.

By comparing the functioning of these computational models to the functioning of cognitive general intelligence, it is easy to identify where they fall short. The gap in the computational models can be bridged by understanding how cognitive general intelligence fills the same gap and replicating those solutions in the computational models.

The first part of the paper, section 3.1., will recap ground covered in the last chapter by briefly re-stating the necessary and minimally sufficient properties needed for general intelligence that satisfies the Grand Ambition. Section, 3.2. will weigh the QTM-AGI, AIXI and LLMs against the necessary and minimally sufficient properties for general intelligence to show that they fall short as examples of AGIs because of a conceptual flaw. Given that these champions for computations have been deemed inadequate, section 3.3 of the chapter will examine if computation is indeed an apt path to general intelligence that is functionally comparable to a mind. The section will show that the mind and computation have enough in common to conclude that the mind is in part performing computations. However, the mind does a bit more than simple computations to realize general intelligence. It leverages processes to ensure computational self-sufficiency. In the last section 3.4., the chapter will propose and defend the mimicry of the functional roles in the mind that support computational self-sufficiency to realize more accurate models of artificial general intelligence.

3.1. Necessary and minimally sufficient properties for AGI.

The Grand Ambition of AI research is to realize general intelligence in machines that is functionally comparable to human general intelligence. From the Grand Ambition, the previous chapter identified the necessary and minimally sufficient properties that will qualify an artefact as possessing general intelligence. Defining the necessary and minimally sufficient properties of an AGI is useful because it creates a well-defined target for the research field to aspire towards, but also a benchmark against which AGI researchers can weigh intelligent artefacts for conformity with the Grand Ambition.

The necessary property for an AGI is problem solving competencies that are general in the sense that the artefact can self-acquire solutions to familiar and newly contextualized problems in its native environment. These solutions are output in the form of actions or behaviour. The artefact's problem-solving competencies should be functionally comparable to human intelligence in a way that is not strictly defined by its capacity or capability.

The necessary property defines the desired extrinsic-facing traits of an AGI. The extrinsic facing traits are the anticipated behavioural features of AGI. They are extrinsic facing in the sense that they are outwardly directed. They define the AGIs output. The extrinsic facing traits are desired because they represent the usefulness of an AGI. The motivation of AGI research is to produce artefacts with cross domain problem solving efficacy that have the capacity to apply their learned knowledge towards solving familiar and unfamiliar problems, with a view of leveraging the usefulness of these features. It is in this sense that they are necessary. They are the definitive property that motivates the research field. To satisfy the motivations of AGI research, the artefact must necessarily interact with its native environment in a manner that allows it to engage intelligently to solve both familiar and newly contextualized problems.

The minimally sufficient properties of an AGI define qualities that are jointly sufficient for an AGI to have functional comparability with human (and other organic) general intelligence. The minimally sufficient properties are intrinsic traits. They are intrinsic in two ways. The most obvious being that they are inwardly manifested. The traits define the AGI's internal functioning. Secondly, they are intrinsic in the sense that they are essential in realizing the type of extrinsic facing traits sought for the Grand Ambition. To satisfy the Grand Ambition's demand for functional comparability with human general intelligence, the extrinsic facing property, i.e., general problem solving, must be realized through –

1. **Access to meaningful inputs that lead to appropriate output.** The artefact must have the ability to receive inputs from the environment that define familiar or unfamiliar problems, and output solutions to the problems.

2. **The input and output must bracket processes that are a cogent and liable transition from input to output.** The successful outputs must be systematically derived from the inputs. The systems that lead from input to output must be appropriate, reliable, and creditable with the success of the output.

3. **The system described in (1) and (2) must be self-initiated and self-directed by the artefact to solve problems.** The artefact, as an agent, must be credited with initiating and owning the successful input to output streams.

The minimally sufficient qualities are minimal in the sense that they are the absolute minimum properties to qualify an artefact as intelligent in a manner that is functionally equivalent to human (cognitive) general intelligence. Together with the extrinsic facing traits as necessary properties, the three jointly sufficient intrinsic traits comprise the necessary and minimally sufficient properties for the type of general intelligence that will satisfy the Grand Ambition.

As minimally definitive of general intelligence, the properties can be adopted to judge artefactual models of intelligence to determine whether they qualify as instances of general intelligence.

It has been proposed that Deutsch's Quantum Turing Machine (Deutsch, 1985), Marcus Hutter's AIXI machine (Hutter, 2000), and given a few improvements in computational power and algorithmic efficiency, popular LLMs (Aschenbrenner, 2024) are models of computational general intelligence. By using the Grand Ambition's necessary and sufficiency properties as checklist, it can be determined whether these theoretical models qualify as instances of the type of general intelligence that satisfies the Grand Ambition.

3.2. Are the QTM, the AIXI and LLMs examples of AGIs in the Grand Ambition Mould?

3.2.1. The necessary properties for an AGI MVP.

While a QTM-AGI, Hutter's AIXI and popular LLMs employ different processes, they have all been designed to address general problems in some form. The QTM-AGI uses quantum processes to represent complex worldly input and supercharge their processing in a Turning mechanism to realize solutions to general problems it has been tasked to solve. Hutter's AIXI utilizes a combination of Solomonoff induction with sequential decision theory to generate its solutions to different problems. While popular LLM models use advanced transformer models to generate human like responses to various tasks.

It should be noted that the QTM-AGI and AIXI are both theoretical formulations, while popular LLMs are real models, that are only projected to satisfy AGI goals given certain improvements. None of these models are real examples of AGI today. Indeed it is likely the case that realizing a practical AGI version of these artefacts will run into physical limitations.

If we are to ignore the practical constraints to building a physical version of these artefacts, the QTM-AGI, the AIXI and the proposed future LLMs functioning as predicted will all meet the necessary condition for general intelligence. The necessary property for an AGI is its extrinsic facing traits. The AGI should necessarily output behaviour or actions that represent solutions to familiar or unfamiliar problems in its native environment. Provided that they have been granted access to suitable resources, in being general computing devices, these artefacts will exhibit the necessary property for an AGI. They will be able to confront and solve both familiar and unfamiliar problems that they encounter. The solutions they output in the form of behaviour or actions should be human-like because they are the result of algorithmic manipulations in that have been designed or fine-tuned by humans to meet human ends.

3.2.2. The jointly sufficient properties for an AGI MVP.

All three of these contraptions have the necessary extrinsic facing traits for an AGI MVP. They are able to solve familiar and newly contextualized problems in their native environments.

However, the Grand Ambition mandates that the necessary extrinsic facing traits must be realized through jointly sufficient intrinsic traits if they are to be deemed to have functional comparability with human (cognitive) general intelligence. The intrinsic traits are the three properties that together are minimally sufficient for the type of general intelligence that AGI research seeks.

i. Access to meaningful input that leads to appropriate output.

If functioning optimally, the three models will derive appropriate output from accessed input that has imputed meaning. The first jointly sufficient condition for general intelligence is that the desired output in the form of behaviour and actions must be derived from accessed input that defines the problem that is to be solved. The engineers of the artefacts have managed to propose designs with successful general input-output streams. The QTM-AGI, the AIXI and future LLMs will employ different processes, but all their outputs are systematically derived from input that they access from their respective environment. Input is information from the world that defines the problem or problems that are to be solved.[12] This information is processed in keeping with the models' operative structures to realize an output that represents a solution to the input-defined problem. Success in all these cases will be defined by realizing an output that represents a rational response to an input-defined problem when judged by a human observer.

ii. The input and output must bracket processes that are a cogent and liable transition from input to output.

The QTM-AGI, the AIXI and LLMs will have appropriate output generated from the manipulation of problem defining input. While employing different processes, these artefacts are in conformity with the second jointly sufficient condition for general intelligence in an AGI. The second demand on the AGI MVP checklist is that the output of the machine must be realized via appropriate problem-solving processes. An appropriate problem-solving process is reliable, adjusted to fit context, and can be credited with success for generating the correct output. An appropriately designed QTM-AGI will use quantum processes to represent the complexity of the world and supercharge the processing power of a Turing machine to manipulate the represented content. If the QTM-AGI is functioning optimally, it will manipulate the input by way of appropriate algorithms that will realize relevant output. The AIXI uses a combination of Solomonoff induction with sequential decision theory to systematically select the right functions that will generate an appropriate output from the input. The advanced LLMs use advanced statistical models to learn and predict patterns in tokenized data. This approach has been demonstrated to be quite efficient at generating output that is like human output at

[12] The world encompasses both internal and external states. The artefact itself is part of the world.

various tasks. Despite employing different approaches, if functioning as predicted, all contraptions will realize processes that are reliable, are appropriate for context, and can be credited with generating the correct output.

iii. The system defined in (1) and (2) must be self-initiated and self-directed by the artefact to solve problems.

While the QTM-AGI, the AIXI and the proposed advanced LLMs will meet the first two sufficiency checks for general intelligence in the AGI MVP checklist, they all fall short of the third condition. The third condition on the checklist mandates that the processes or functions that lead from accessing input to the output, must be credited to the artefact itself. This means that the artefact must have the autonomy that is creditable with initiating and exercising enough control over the successful input to output stream to own the solutions that it outputs. This mandate can be understood as a demand for self-sufficiency.

The phrase self-sufficiency means an ability to maintain or sustain oneself / itself without outside help. From this description, one can derive both a negative and a positive dictate for self-sufficiency. The negative dictate defines an absolute prohibition on any extrinsic agency in processes that are defined as self-sufficient. The positive dictate establishes an intrinsic agency, which is credited with the self-sufficiency.

The third condition on the checklist for general intelligence in the Grand Ambition is that the functions or processes that lead from accessing input to the output should conform to both the negative and positive dictates for self-sufficiency. They must be initiated or directed without outside agency, and the process/es must be credited to the artefact itself.

The QTM-AGI, the AIXI the proposed LLMs fail to meet the negative dictate for self-sufficiency. They rely on programs that have been designed and fine-tuned by extrinsic agents to execute problem solving in a manner that satisfies the extrinsic agents. The role of the extrinsic agents includes prescribing the type of inputs and the rationality for how they should be processed. In being a Turing machine, the QTM-AGI accesses inputs in a form prescribed by an extrinsic agent and manipulates the input according to extrinsically designed instructions that are installed into the machine. The AIXI relies on various sets of extrinsically formulated instructions (programs) that are pre-installed into the machine and made searchable. The machine then automates the selection of the appropriate set of instructions that will realize the optimum solution. The LLMs use advanced statistical models that are fined tuned by extrinsic agents to identify and predict patterns that are useful for certain tasks. The QTM-AGI, the AIXI and the LLMs can only solve the problems they have been armed to solve through extrinsic processing instructions, or extrinsically fine-tuned statistical models. In all these cases, extrinsic agents bear an inordinate amount of credit for problem solving. They have formulated or finetuned the problem-solving functions and installed them in the machine. These approaches violate the negative dictate for self-sufficiency which puts a prohibition on extrinsic agency in self-creditable problem solving. All these artefacts' problem-solving capability is creditable in large part to the programmers.

All these artefacts also violate the positive dictate for self-sufficiency. The positive dictate for computational self-sufficiency is a demand for a 'autonomy that can be credited with problem solving. These artefacts do not have clear intrinsic processes or states that can be identified as an inner-self that can be credited with problem-solving. They are more akin to automatons, in that they perform a range of functions according to predetermined set of coded instructions or adjustments, rather than possessing states or events that perform ratiocination that is creditable with problem solving. Orseau, et al. suggest that we should adopt the terms 'device' and 'agents' to mark this distinction within computational theory (Orseau, et al., 2018).

None of the artefacts have the type of self-sufficiency that can be wholly credited with generating their output. As a result of falling short of the third condition in the checklist for general intelligence, the QTM-AGI, the AIXI and the advanced LLMs do not have the sufficient conditions for general intelligence in the Grand Ambition mould. The three artefacts are very clever contraptions, but they are not the type of general intelligence that the Grand Ambition seeks.

3.2.3. Output-centric approaches to AGI and the AGI MVP checklist.

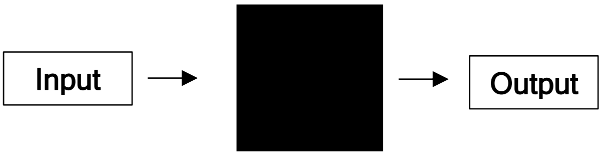
The QTM-AGI, the AIXI and advanced LLMs are output-centric attempts at realizing an AGI. An output-centric approach to AGI places emphasis on building models that reproduce human-like or human-desirable input to output patterns, without a commitment to realizing any functional processes or states that yield the human-like intelligent output. The QTM-AGI, AIXI and LLMs employ different processes, none of which have an obvious homology with principles that yield cognitive general intelligence. The engineers of the systems are labouring under a motivation to realize any system that most appropriately yields output in the form of behaviour and actions that is comparable to human behaviour and actions when addressing general problems in a native environment.

The output-centric attempts towards realizing artificial general intelligence are conceptually flawed. They are destined to violate the third demand for general intelligence in the AGI MVP checklist. As a reminder, the third demand on the AGI MVP checklist is a demand that the artefact, itself, is credited with initiating and directing the processes that yield the solutions to problems. It is a requirement for self-sufficiency in problem solving.

An output-centric model is doomed to fall short of self-sufficiency because it relies on extrinsic agency to carry out problem-solving. To explain why output-centric computational approaches are challenged by the demand for self-sufficiency, consider the brief explanations of general intelligence below.

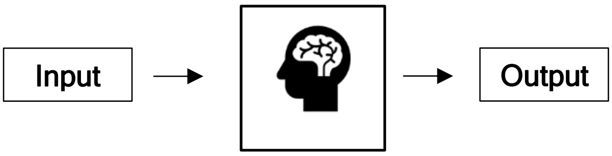
Assume for the objective of explanation that the schematic illustration below illustrates non-descript **general intelligence** (image 3.1). Input from the environment that represent different problems are fed into a black box which yields outputs that represent optimized solutions to the relevant problems in the environment.

Image 3.1- General intelligence schematic.



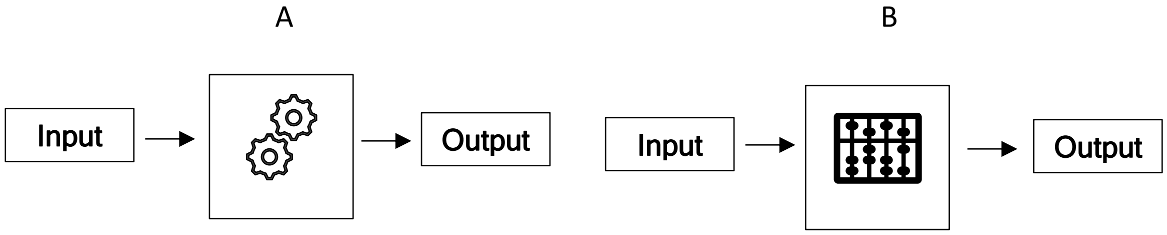
For **human intelligence** (image 3.2.), the box between input and output is filled with specific self-initiated cognitive processes (represented in the schematic below with a human head with a brain). Apart from the specific cognitive processes that are associated with the brain, the human general intelligence works the same as above. The humans receive inputs from the environment that represent different problems in the environment, they process it through processes in the box (in this case cognitive processes) and return outputs that are optimized solutions to the problems in the environment.

Image 3.2 - Human general intelligence schematic.



The **output-centric approaches** (image 3.3.) to AGI approach the black box as an opportunity to be creative with the contents of the box between input and output. The aim is to match the input to output patterns in the human intelligence schematic above without a commitment to the specific cognitive process in the box.

Image 3.3 - Output-centric model schematics.



Schematic A and B (above) represent two different output-centric approaches to AGI. As an example, one might assign the schematic A to represent Hutter's AIXI and schematic B to represent an advanced LLM. The processing icons (the gears in process A, and the abacus in process B) are contrasted against the human head in the human intelligence schematic (and against each other) to underline that they are functionally different processes.

The output-centrist approaches the AGI problem with an optimistic attitude that there are several types of processes that might yield comparable human-like outputs from inputs in the environment and all of them will be sufficient to be deemed general intelligence. In this sense, the output-centrist is a functionalist because they believe that the output that conforms to general intelligence is realizable through multiple processes.

However, even if the output-centrist's argument that there are several processes capable of realizing human-like input to output patterns was granted, the output-centrist will still fall short of realizing general intelligence. This is because in approaching the box between input and output as an opportunity to be creative with the processes that lead from input to output, the output-centrist must bear the onus to ensure that whatever processes they chose to fill the box must be carefully calibrated, designed, fine-tuned or programmed to produce the same results as human processes.

For example, the AIXI's stored programs must be purposefully designed and arranged to ensure that they convert input to output in a way that matches human input - output patterns, and the QTM-AGI's Turing table of rules must be programmed to process the input to match human output. The LLMs must have their algorithms trained and fine-tuned to ensure that their output matches the desired human output. Ensuring that the output-centrist's processes of choice are calibrated to match human output, entails a designer working backwards from human comparable outputs to purposefully design the alternative system to effect the desired output. The output-centric computational models, therefore, have problem solving efficacy that can only be creditable to their programmers or designers. Their problem-solving is creditable to an extrinsic agent. As a result, they cannot be credited with self-sufficiency in problem solving. This is a violation of the third sufficiency check for an AGI MVP.

Even though output-centric approaches to AGI are a conceptually flawed, they are the primary computational approach in AGI research. This glaring conceptual flaw in mainstream output-centric models raises the question as to whether any computational based models can account for the functioning of general intelligence. In trying to get to the bottom of this question, it is worth exploring whether computation is indeed an adequate explanation of the mind.

3.3. Computation, the mind, and AGI.

An effective way of exploring whether computation can account for the mind is to work backwards from the mind to determine whether it can be explained by computation. The only example, we currently have of general intelligence is organic general intelligence such as exhibited by cognitive agents. Many believe that the quintessential example of cognitive intelligence is the human mind. This blatant anthropocentric bias is most likely in place because as humans, we are most familiar with and have the most access to our own minds for interrogation. To determine whether computation can account for the mind, it is worth exploring whether the mind is doing computation. Such an exploration must begin with a definition for computation.

3.3.1. Computation.

Computing is the process through which a system uses algorithmic processes to simulate, describe, or project the transformation of information that represents states of affair or knowledge from the world. The input of a computing system is structural data that describes a qualitative state in the world at T1. The structural data is input into the system as represented information in a form that is manipulatable by the system. The process of computing then systematically manipulates the represented structural information under appropriate rules to realize a form that either simulates changes, describes, projects, or engages the transformation in the input at T2.[14]

Two strict requirements for computing stand out-

i. Information from the world is represented in some manipulatable form.

[14] Questions have arisen about whether mathematical computations align with a description that defines computation as being about the world. The answer is yes. Mathematics are methods and principles used to understand complex aspects of the world (or reality) by manipulating quantified structures and relationships to predict future events, describe realities, and/or solve problems.

ii. Represented information is manipulated under formal system structures.

i. Represented worldly input.

The first requirement for computation is that worldly information must be represented in a form that the system can manipulate. This represented information is often referred to as symbols. Symbols are manipulatable primitive data forms that represent information from the world by bearing semantic properties. It is useful to think of semantic properties as intentional linkages. A computational symbol represents an object or state of affair in the real world by being linked to the object or state of affair. The symbol has meaning in the sense that they represent the world as being a certain way. Even though symbols have semantic properties, these properties do not determine the form of the symbol. The form of a symbol is determined exclusively by fit into the formal structure of the system. The reason for this is explained by the second requirement for computation.

ii. Formal manipulation.

The second requirement for computation is that the represented information from the world, in the form of symbols, must be manipulatable. This is a demand for the computational system to have syntactic structure. The symbol must have a form that can be manipulated by the formal movements of the system. A system that is apt for computing has formal movements and structure that can be explained by or manipulated to follow firm deductive principles. A system that has internal operations that adhere to defining deductive principles is said to be expressing a formal language. A computational symbol is a native form of a formal language. By nature of its form, it is a subject to the principles that govern the syntactic movement and structure of the system. It is in this sense that a symbol is manipulatable by the system. The manipulation of the symbols is implemented under rules that are embodied by the formal language. If the syntactic rules are well formulated, they should cohere with the structure and dynamics in the world that they represent. All computation is the formal manipulation of syntactic forms. These forms bear semantic properties, but the computation is affected only in virtue of syntactic features.

There is little debate that formal systems like the QTM-AGI and AIXI meet the two essential requirements for computation. The QTM-AGI utilizes quantum data types to represent complexity or knowledge from the world (symbols) and manipulates these data types according to user-specified formal instructions (syntactic structure). Both the symbols and syntax in the QTM-AGI are designed to align with real-world events and dynamics that need to be computed. Similarly, the AIXI is equipped with a set of pre-formulated user instructions, presumably comprehensive enough to cover the full range of human behaviour. These instructions are encoded in a formal language (syntactic structure), with the input from the world is represented as native data types (symbols) in that language. Depending on the input, the AIXI evaluates all the programs to determine which one offers the best ratio of economy in terms of computational resources to reward, and the system choses the most optimal program to manipulate the symbols representing the input.

Although less obvious than with the QTM-AGI and AIXI, large language models (LLMs) also meet the two requirements for computation. LLMs operate using neural networks. In neural networks, input from the world is represented as distributed weights and activations rather than localized tokens (symbols). This approach marks a significant departure from the other two systems. However, it still satisfies the first requirement of computation. The distributed representations are then processed according to formal statistical rules that treat the distributed weights and activations as native forms of the system itself (syntactic structure). This aligns with the second requirement for computation. Like the QTM-AGI and AIXI, LLMs rely purely on syntactic processing, without regard to semantic properties. However, while they are formal machines, they are designed so that their internal manipulation and syntactic structures correspond to real-world events.

3.3.2. Is the mind a computer?

The next question to resolve is whether the mind is a computer. The mind is the element of an organic cognitive agent through which the agent has an awareness of information from the world and effects processing of this information.

For most AGI researchers, there are enough parallels between mind and computation to sustain an analogy. For example, the mind does seem to receive information from the world, represent it in some mental form, processes it through internal movements and manipulations to produce an output that represents transformations, projections, or knowledge of the world.

A computational theory of mind (CTM) has been the orthodox view of the mind in cognitive science and A.I. research. As stated earlier, the initial adoption of CTM came off the back of Turing's description of a universal Turing machine (Turing, 1936). Turing described a universal Turing machine (UTM) that was a programmable general-purpose computer that could mimic any other Turing machine. Turing's UTM inspired Warren McCulloch and Walter Pitts to make comparisons between Turing's UTM principles and the mind (McCulloch & Pitts, 1943).

The view that the mind was doing some form of Turing styled computations inspired the initial stirrings of AI research and came to be known as Classical Computational Theory of Mind (CCTM). The CCTM holds that mental processes such as reasoning and decision making in problem solving are computations that are not unlike the computations performed by a Turing machine (Rescorla, 2015). It is important to keep in mind that the CCTM is a family of views.

The exact nature in which mental processes are like Turing processes varies among different CCTM theories, none of which support a strict like for like comparison between a Turing Machine and the mind (Rescorla, 2015).

The most prominent CCTM theory is a Representational Theory of Mind (RTM). The RTM's most ardent advocate is Jerry Fodor (Fodor, 1975). The RTM postulates that the mind represents information through primitive symbols that can be combined into complex symbols. Mental activity works by manipulating these compositional symbols through Turing styled computations. Fodor describes mental symbols as expressions in a formal language of thought that he calls Mentalese. Fodor takes a realist view on mentalese symbols. He believes that mentalese symbols are a real expression in a natural language like logic in a literal sense (Fodor, 1980).

Others have taken a more pluralistic view of the symbols in the mind. They are open to and have posited different types of mental representations such images, maps, diagrams, and other non-propositional forms (Johnson-Laird, 2004, p. 187), (McDermott, 2001, p. 69). A pluralistic view on the nature of mentalese symbols does not affect the fundamental principles of RTM. The RTM seems to work just as well if mental computation is envisioned to operate over a pluralistic interpretation of representations (Pinker, 2005, p. 7), (Sloman, 1978, pp. 144-176). The fundamental principles of computation only require that information is represented in some manipulatable form, and that the information is manipulated by virtue of this form under appropriate rules.

The idea that the mind is a computer was brought into question by the emergence and maturing of connectionist explanations of the mind. Connectionism threatened to tear down the orthodoxy that was the CCTM (& RTM) as an explanation of the functioning of the mind. The main handicap of the CCTM (& RTM) are that they are an ill fit with the physical structure of the brain.

Connectionism is inspired by neurophysiology. Rather than approach computation in the style of a Turing machine, connectionism employs a collection of interconnected nodes arranged in a neural network. Neural networks tend to have three types of nodes. They are called the input nodes, the hidden nodes, and the output nodes. The information to be computed is fed into the neural network through the input nodes. This information is relayed to distributed hidden nodes where it is manipulated through weighted connections between the nodes. Finally, the information is passed to the output node as a computed output. This type of neural net processing appears to be a better fit with the physical structures of the mind.

Initially, it appeared that connectionism and the CCTM (& RTM) were radically different approaches to computing. They were framed as diametrically opposed theories. The irreconcilable difference seemed to be that CCTM (& RTM) is inextricably tied to computing by rule-governed symbol manipulation while neural networks do not seem to be manipulating any obvious symbols. Neural networks appear to be performing non-symbolic movements. Because of this critical difference, Eliminativist Connectionists believed that connectionism should replace Turing styled formalism as an explanation of the mind (Churchland, 1989); (Rumelhart & McClelland, 1986). Connectionism was deemed to be more faithful to the physical structure of the mind.

Adopting connectionism as an explanation of mind threatened to bring into question whether the mind qualifies as a computer under the *de rigueur* description of computing as symbol manipulation. A fundamental requirement for computing is the systematic manipulation of symbols that represent information from the world. If a system is doing non-symbolic processing, can it be considered a computer?

The instinct to position CCTM (& RTM) and connectionism as mutually exclusive explanations of the mind is unwarranted. Despite the two interpretations appearing to be radically different approaches, there are enough similarities to abstract common computational principles. The suggested irreconcilable issue is that connectionism does not seem to utilize symbols in its processing. This appears to violate the critical definition of computation as the processing of represented worldly information by virtue of the manipulation of its representational form. However, connectionism can only be described as non-symbolic if one adopts a rigid view of represented information as localized symbols. In neural networks, information is represented through distributed weights and activations. Neural networks might not represent information in the form of the localized symbols that are most commonly envisioned in the CCTM and RTM, but neural networks do represent and manipulate distributed symbolic information by virtue of its representational form.

It should be noted that the most popular notion of symbols in the CCTM family of theories is inspired by Turing's take on computations. In describing the principles that govern the Turing machine, Turing made only two demands of symbols. The first is that there are a finite number of primitive symbols that can be represented in computation. The second is that the symbols can be combined into complex structures. While his description of a Turing Machine did utilize symbols represented in a localized version, there is no evidence to suggest that localized symbols are fundamental to his theory.

Without a demand for locality of representation, neural networks satisfy both demands that Turing made of symbols. Distributed representations can express a finite number of primitive symbols, and these symbols can be combined into complex structures. In neural networks, these distributed symbols are represented and manipulated by rules expressed as weights between connections. These rules manipulate the distributed representations by virtue of their form. Connectionism should be understood as a type of computation that manipulates non-localized represented information.

Connectionism and the CCTM family of theories have enough in common that we can group them as related theories under the broader CTM. Some have even gone as far as classifying them as amalgamable parts of the same theory. Implementationist connectionism argues that a true explanation of the mind will include both classic computational models and neural networks operating at different levels of explanation (Marcus, 2001), (Smolensky, 1988). There is good reason to suggest that implementationist connectionism might be right. Afterall, all the artificial neural networks that have been constructed to date have been built on classical computers; and in the same vein, classical computing models can be realized in neural networks (Rescorla, 2015). The two schools of thoughts have clear amalgamable overlaps.

From a common-sense point of view, the evidence suggests that there is enough in the CTM to sustain a description of the mind as doing computation. Folk psychology states that mind hosts mental content about the world that is presented in some intrinsically accessible form, and that this mental content is subject to manipulation through mental processes like thinking, calculating, projecting etc. The CTM asks us to think of mental content as forms of mental activity with representational properties. The represented properties, whether expressed on distributed or localized structures, are processed through the systematic manipulation of the representational forms. Both the CCTM (& RTM) and connectionist views of the mind qualify as CTM models.

The strength of the CTM as an explanation of the mind is that it conforms to folk psychology. From a folk psychology perspective, the mind does appear to be doing computations. There appears to be enough to sustain an analogy between computation and some aspects of the functioning of the mind.

This raises the question as to why computation has, as of yet, not yielded the type of general intelligence that the mind exhibits. Where is it that computational models like the QTM-AGI, AIXI and the LLMs are falling short in matching the mind’s capacity for general intelligence? While the mind does appear to be doing something akin to computation in representing worldly input and manipulating the representational forms to inform output, there appears to be something critical missing for computational models to bridge the gap to mindlike general intelligence.

3.3.3. The gap between computation and the mind.

The mind in being the widely accepted quintessential example of general intelligence should comfortably meet the necessary and sufficient properties for AGI in the Grand Ambition mould. Indeed, the human mind is the primary muse for the AGI agenda, and an example drawn from the human mind was employed in the last chapter to extract and defend the intrinsic properties that populate the checklist for general intelligence in the grand ambition.[19] Therefore, a good approach to examine where computational models like the QTM-AGI, AIXI and LLMs are falling short of general intelligence is to examine where the mind succeeds, but the computational models fail.

As a reminder, the sufficiency checklist mandates that an artefact with just enough features to be labelled as possessing general intelligence, must possess problem-solving approaches to familiar and unfamiliar problems that involve the following -

1. Access to meaningful inputs that lead to appropriate output. The artefact must have the ability to receive inputs from the environment that define familiar or unfamiliar problems and output solutions to the problems.
2. The input and output must bracket processes that are a cogent and liable transition from input to output. The successful outputs must be systematically derived from the inputs. The systems that lead from input to output must be appropriate, reliable, and creditable with the success of the output.
3. The steps in (1) and (2) must be self-initiated and self-directed by the artefact to solve problems. The artefact, must have the autonomy be credited with initiating and owning the successful input to output streams.

A keen observer would correctly identify that the first two jointly sufficient conditions (in bold) for general intelligence define computation. Computation is the process through which a system converts accessed input that represents states of affair or knowledge from the world to output (1) by way of algorithmic processes that simulate, describe, project, or engage the transformation of the symbolic information (2).

However, it was noted that output-centric computational models fall short of the third sufficiency condition for general intelligence. The third condition is that the computational steps (1) and (2) must be creditable to the model itself. This is a demand that the computation defined in (1) and (2) is carried out in a manner that is self-sufficient. Computational self-sufficiency

[19] See chapter 2 section 2.3.

mandates a strict prohibition on extrinsic agency in problems solving, and a requirement for self-owned processes, states, or functions that can be credited with problem-solving.

The most common computational approaches are output-centric models, which lack the capacity for computational self-sufficiency. This is because they are reliant on extrinsic agency to solve problems. The mind is not handicapped with this deficiency. The mind is computationally self-sufficient. It manages both the computational steps in (1) and (2) without extrinsic agency, and in a manner that is creditable to a self that is emergent from its internal processes.

It is evident that computation does explain some functional aspects of general intelligence. But it is not the whole story. Computational models need to be ameliorated with processes that support computational self-sufficiency to bring them closer to the general intelligence that is interesting to the Grand Ambition. Lessons can be learnt from how the mind supports computational self-sufficiency.

3.3.4. The two problems faced by output-centric computation.

The mind in cognitive general intelligence is capable of effecting problem solving in its native environment self-sufficiently. As stated earlier, computation involves accessing inputs from the environment that have meaning which define familiar or unfamiliar problems. And manipulating the represented inputs in cogent, reliable and systematic way to yield a desired output. Self-sufficiency implies that both these steps are carried out without reliance on extrinsic agency and in a way that is creditable to intrinsic agency.

The mind satisfies both these demands while output-centric models, such as the QTM-AGI, AIXI and LLMs fail the test for self-sufficiency. These models fail both the negative dictate of self-sufficiency, a strict prohibition on extrinsic agency, and the positive dictate for intrinsic states and process that are credited with autonomy. This is because they rely on extrinsic agents to prescribe how problems are defined and processed.

Two handicaps stand out for why output-centric systems will find self-sufficiency challenging.

i. The grounding problem.

Consider the first requirement for computational self-sufficiency. The first requirement is the ability to access inputs that have meaning which define a problem to be solved. Output-centric systems like QTM-AGI, AIXI, and LLMs rely on external agents to assign meaning to both the problems and the inputs that define the problems. From the system's perspective, the problems it solves are meaningless. The QTM-AGI, AIXI, and LLMs are simply executing externally prescribed instructions or statistical processes. In these systems, there is no need for the defining input to hold any intrinsic meaning within the system itself. This problem relates to the symbol grounding problem in philosophy and cognitive science. The symbol grounding problem refers to the challenge of connecting computational items in a computational system to their real-world referents without external intervention or interpretation (Harnad, 1990).

The problem is as follows. Computation is symbol manipulation under syntactic rules that manages in some way to maintain conformity with affairs in the world that the computation aims to simulate, represent, or project. Symbols are a critical feature of the coherence between computation and the world because they are a bridge between the world and the formalism of the system.[20] Symbols are syntactically manipulatable forms that bear semantic properties. The semantic properties are representational content that represents the world as being a certain way. For the symbols to be manipulatable, they must be expressed in a form that adheres to the syntax of the formal structure of their system. The manipulation of the symbols is implemented through inferential syntactic rules that are expressible within the formal structure of the system itself. In other words, even though symbols bear semantic properties, they are expressed as syntactic objects and manipulated under syntactic rules. If the syntactic rules are well formulated, they should cohere with the semantic structure and events in the world that they represent.

This raises a critical question; How does the syntactic forms and rules of computation find coherence with the world? All computation is the rule-based manipulation of formal syntactic forms. While the syntactic forms have semantic properties, the computation is affected only in virtue of syntactic features. If computation is blind to semantic properties, how does it under the execution of its syntactic rules, manage to manipulate computational items in such a way that there is a rationality in their manipulation that is true to the phenomenon and dynamics that they represent in the world? Even more fundamentally, how does the syntax in the mind acquire semantic properties? How is worldly meaning imputed on what appears to be closed syntactic structure?

The answer to the problem in output-centric systems is an extrinsic agent. Non-cognitive computation owes its internal coherence to the world and its dynamics to outside agency. Artefactual computation acquires meaning and coherence from programmers and engineers. The semantic properties of symbols in artefactual computers are in the mind of the programmers (and the users), they are simply assigned, imputed, or prescribed to symbols. The programmers of computers either discover structural principles in a system that coheres

[20] Read the word ‘symbol’ as any primitive data form, whether localized, distributed, projected, or integrated in the physical structures of the system, which represents information from the world by bearing semantic properties.

with their intentions for computing or manipulate a systems' syntactic structure to cohere with their intentions. Computers are either purposefully interpreted or designed and programmed to manipulate symbols in a manner that reflects the dynamics in the world that the system represents. The CTD principle states that if the programmer is particularly clever and has unlimited resources, they can manipulate syntactic structure to ensure manipulations that will closely cohere with any natural system that they would like to model (Deutsch, 1985). For artefactual systems, syntactic and semantic coherence is a purposeful endeavour realized by a programmer.

For many computational cognitivists, the fact that syntactic structure in artefactual computers can be designed to cohere with semantic structure is further evidence for the CTM. It is apparent that the minds processing has intentionality that is about the world. Given that computers can be programmed effectively to manipulate syntactic forms that have been imputed with semantic properties, it can be abductively concluded that the mind processes information under the same principles. Fodor takes this line of thought to argue that CTM is the only explanation of the mind that can account for how mental activity tracks semantic properties in a coherent way (Fodor, 1987, pp. 18-21). His argument is as follows; Mental activity coheres with semantic properties. Computer science has shown us that syntactic manipulations can be done in such a way that they track semantic properties. If we take the mind to be a syntactic machine, then we can explain why the syntactic manipulation in the mind coheres with semantic properties. Therefore, we can explain how mechanical rationality can be possible.

However, the mind does not rely on an extrinsic agent to ground mental symbols. Fodor's abductive reasoning is an argument for the CTM, but it does not entirely resolve the problem of how the mind grounds symbols without recourse to an extrinsic agent as is the case with artefactual computing. If computational cognitivists are right and the mind is a syntactic engine, then it must follow that the mind is manipulating mental symbols that are defined by their syntactic properties rather than their semantic properties. Symbols are primitive and complex items individuated by their formal structure and computation of these items is determined by syntactic rules operating on the formal structure rather than on the semantic properties. Both classical computationlaist and connectionist endorse this generalized view of formal processing. They believe that computation in its formal movement is blind to semantic properties. If CTM's interpretation of the mind as performing computation is to be sustained, there must be an explanation of how meaning related to the world finds expression in the mind's syntactic machine.

Artefactual systems rely on external agency to link syntax to semantics. It is almost certain that no computational cognitivist would suggest that mind has the same extrinsic purposeful programming as an artefactual computer. If the mind is a syntactic engine, how does the mind's syntactic manipulation cohere and remain true to the rationality in the world?

ii. The problem of self-control.

The second demand for computational self-sufficiency is a demand for internal states or processes that can be credited for realizing intelligent behaviour. This idea found an early articulation in Book 3 of Aristotle's Nicomachean Ethics, where Aristotle defined behaviour as actions that have their origin within the agent (Aristotle, 2000). While the way agents initiate action is still a matter of much debate, there is a consensus that actions credited to the agent are goal directed. The dominant thought in philosophy of action is that this involves the agent owning an intention to perform the action. In lay terms, self-creditable action denotes a capacity to act intentionally, or in other words, the ability to take intentional action (Glock, 2019), (Kenton, et al., 2023), (Liljeholm, 2021), (Orseau, et al., 2018).

Intentional action is a behaviour driven by an expectation that the behaviour is likely to bring about a desired outcome (Dickinson, 1985). Dung adds that intentional agency is further defined by autonomy and the system's ability to act for its own reasons (Dung, 2024). The Standard Theory of Action describes in causal terms what constitutes intentional action and therefore reasons that support intentional actions. The theory states that an intentional action is an action that is caused by the agent's right mental state/s, acting on or in response to the right event/s, in the right way (Schlosser, 2019). The right way in this case is an action that is rationalized from the agent's point of view as an appropriate means to approach occurrent events. In other words, the outcome of the action represents an appropriate goal for the agent given the occurrent states and events. Often the right mental states are defined in terms of beliefs and desires (Goldman, 1970), (Davidson, 1980), (Dretske, 1988). Therefore, in the standard theory of action, the reason for actions is quite frequently defined in terms of rational ends that are derived from the agent's mental states such as beliefs and desires interacting with or in response to occurrent events for the agent's end.

Many believe that AIs are not capable of intentional actions (Xabier E. Barandiaran, 2009, p. 382). An argument for this stance might proceed as follows. (1) Human (cognitive) intelligence is exercised through intentional actions. (2) These intentional actions are furnished through mental states. (3) Mental states cannot or should not be attributed to artificial intelligence. Therefore, artificial intelligence cannot have intentional actions. The temptation is to dismiss the first premise on the grounds that it is blatantly anthropometric. However, this anthropometrism can be excused because some factor of human intelligence is often the goal or the measure through which artificial intelligence aspires or is judged. We can see this clearly reflected in the Grand Ambition of AGI which is to realize artefacts with general intelligence that is functionally comparable to human general intelligence. Premise two would not draw any serious objection, therefore the question on whether AIs can take intentional actions is going to turn on premise three, can we or should we attribute mental states that furnish intentional actions to AI?

This thesis subscribes to Dennet's instrumentalist view on agency. The instrumentalist position is best encapsulated in Daniel Dennet's intentional stance (Dennet, 1987, p. 17). Dennet suggests that we should attribute intentional actions to a system to the extent that taking an intentional stance towards the system is useful in explaining and predicting its behaviour (Shoemaker, 1990). According to the intentional stance we should not shy away from defining AIs' representational states in the language of intentional mental states (e.g., beliefs, desires, intentions), if attributing these states to an AI can explain or predict its rational behaviour.

Even from Dennett's instrumentalist perspective, output-centric systems like AIXI, QTM-AGI, and LLMs lack states that directly parallel the mental states associated with intentional actions. The problem is that we cannot identify any internal states within these systems that have intrinsic intentional content. Since these systems cannot ground their own meaning, the states that support problem-solving are assigned or imputed with meaning extrinsically. None of the problem-solving states in these systems have intrinsic meaning that could be compared to intentional states like beliefs, desires, or intentions. These systems are built to solve externally prescribed problems using externally prescribed meaning.

Without a satisfying physical explanation for the mechanics of how the mind support computational self-sufficiency, and therefore why computational models fail, the CTM will continue to be brought into question. For some, the role of a third party in the grounding of artefactual systems is such a radical departure from the self-sufficiency of cognitive processes that the gap could not possibly be bridged. Thinkers like John Searle and Kenneth Sayre argue that this fundamental difference between the mind and computers is sufficient reason to abandon the analogy of the mind as a computer altogether.

Sayre points out that computers do not operate on symbols with intrinsic semantic content. Therefore, pointing to the solely syntactic operations of computers is not useful for understanding how the mind processes symbols with meaning or performs any semantic information processing (Sayre, 1986, p. 123). There are no examples of artefactual systems that have meaning which is independent of outside interpretation. Since computation is defined by strictly formal terms and mental activity is driven by intentional terms, computing by itself cannot explain the mind.

Searle agrees with Sayre's position. Searle believes that our understanding of the concept of computation is that computations are purely formal. They manipulate symbols based on their syntactic structure and not by their meaning. Therefore, CTM could not account for semantics and intentionality (Searle, 1984). If the mind is doing computations, it is not the definitive process. The mind is certainly doing more than just formal manipulations. Searle is not opposed to the idea that computations might define some part of the mental processes. However, he thinks that this a trivial insight because almost any physical system can be described as executing computations. A wall could be said to be performing computations. If we could discern the patterns of molecular movements in a wall, it would surely cohere to some formal structure that can be defined as computation (Searle, 1990).

For Searle, symbol and syntax used in computation is observer (user) relative. Computational symbols do not have meaning that is intrinsic to their nature. Whatever meaning formal symbols possess is imputed, assigned, used, programmed, or interpreted into the physical state that is being computed (Searle, 1990, p. 26). Therefore, if the mind is doing computations, it is the least interesting thing about the mind. The interesting part of the mind is what gives mental symbols meaning. Therefore, what should define the mind is the process by which it attaches meaning and causes symbol manipulation.

3.4. Replicating the self-sufficiency of cognition.

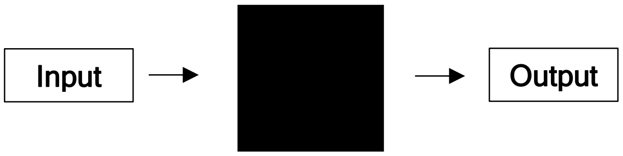
If the processes that support the mind's computational self-sufficiency can be understood, these roles can be distilled into replicable principles to guide the engineering efforts in AGI research. Such an approach will qualify as a process-centric approach to AGI.

As a reminder, a process-centric approach places emphasis on recreating the functional processes or states that yield general intelligence in humans (or cognitive agents). The mimicry approach will aim to distil principles from the functional processes in cognition that support general intelligence and recreate them to realize artefacts with general intelligence.

It is worth noting that unlike the output-centric approach, a process-centric approach is not handicapped by conceptual constraints in meeting the sufficiency conditions for general intelligence. This is demonstrable with the aid of the illustrative schematics that were employed earlier in the chapter.

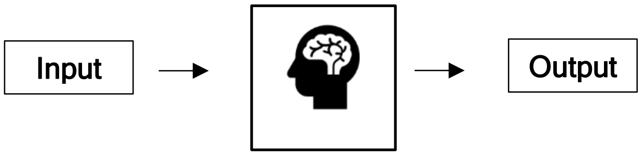
Consider once again that the illustration below (image 3.4) illustrates non-descript general intelligence. Input from the environment that represent different problems is fed into a black box which yields outputs that represent optimized solutions to the relevant problems in the environment.

Image 3.4- General intelligence schematic.



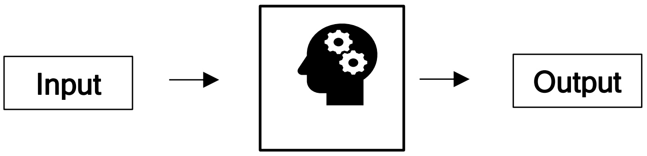
The schematic below (image 3.5) represents human general intelligence. The box between input and output is filled with self-initiated cognitive processes that are represented in the schematic with a brain in a human head. The human intelligence works the same as the schematic for general intelligence, except that instead of a black box, the humans receive inputs from the environment that represent different problems in the environment, and they process it through specific cognitive processes to return outputs that are optimized solutions to the problems in the environment.

Image 3.5- Human general intelligence schematic.



The process-centric approach to AGI (image 3.6) aims to fill the black box with processes that function under the same principles as human intelligence. The processes are represented in the schematic below in the form of a human head with gears. The gears in the schematic represents a material disparity from the cognitive processing in human intelligence (represented by a brain in the human intelligence schematic above). However, the human head in being like the head in the human intelligence schematic signifies that the processes in the box function under the same principles as human intelligence. The process-centric approach to artificial general intelligence places the emphasis of functional comparability to human intelligence on the processes that yield optimized solutions in human cognition.

Image 3.6- Process-centric model schematic.



The process-centrists' focus is on modelling the processes of human intelligence rather than simply recreating human input to output patterns. This is because they believe that the right output would naturally follow from a model that functions under the same principles that govern the specific processes within the human mind. Even though the process-centric engineer is faithfully committed to the specific principles that govern human cognition, they are still functionalist in the sense that they believe that the principles that define the processes in the mind are multiply realizable.

Unlike the output-centrist, the process-centric engineer is not burdened with a need to work backwards by working out human-like output and then attempting to calibrate, program or design structural solutions that will match the same results. This is because the process-centrist is interested in recreating principles that govern human intelligence. If they succeed, their artefact will function under the same principles as human intelligence, even though it might be constituted of dissimilar materials. Two systems that are functionally equivalent will produce the same results given that all other factors remain constant. A successful process-centric system would be as self-sufficient as human intelligence at solving problems.

3.5. Conclusion.

Computational strategies towards AGI have been mainly pursued through output-centric computational models. These include theoretical formulations like the QTM-AGI, the AIXI and most recently LLMs. Output-centric computational models are a conceptual strategy that aims to realize artificial general intelligence by prioritizing the recreation of human-like input to output patterns.

Computation has proven to be an apt strategy in recreating human-like input to output patterns over narrow domains. Computation as a concept is accommodative of different representational and algorithmic processes, requiring only that they drive the requisite input to the desired output. In well defined, narrow tasks, computation has proven remarkably successful at recreating human-like input to output patterns. This is because the relative lack of complexities in narrow tasks makes it easy to calibrate or program representational systems and their structures to match the information and dynamics that they represent in the real world.

However, as an approach to realizing the type of general intelligence that the Grand Ambition seeks, output-centric computational approaches are conceptually flawed. The Grand Ambition is to realize artefacts with general intelligence that is functionally comparable to human general intelligence.

While there is enough in mental processing to sustain an analogy with computation, it is evident that mere computation is not the full story. The mind manages these processes without recourse to extrinsic agency and in a manner that is creditable to a self that is emergent from its internal processes. It is this gap that motivates Dreyfuss in stating that computational models have only covered the distance of a chair’s height to the moon in their attempts to realize the Grand Ambition (Dreyfus, 1978, p. 12).

The best approach to cover the distance towards AGI is to understand how the mind manages to be computationally self-sufficient and to extract replicable principles from these processes that can be used to ameliorate computational models with computational self-sufficiency.

This approach places the emphasis of AGI research on modelling the processes of human intelligence rather than simply recreating human input to output patterns. If done correctly, the process-centric engineer will not be burdened with a need to work backwards to calibrate their model’s computational structures to produce the right output. This is because the process-centrist is primarily interested in building an artefact that functions under the same principles as human intelligence. If they succeed, their artefact will have functional equivalence with human intelligence, and therefore will produce comparable output given that all other factors remain equal.

The next chapter will explore how the mind satisfies the negative dictate for computational self-sufficiency. The negative dictate is a strict prohibition on extrinsic agency in computational problem solving. The chapter will examine why mainstream output-centric models fall short of this dictate. This is because they are reliant on extrinsic agents for symbol grounding. The chapter will then extract a replicable principle from cognitive self-sufficiency that can be used to support computational self-grounding in AGI.

Chapter 4: The Chinese Room and the Analog Principle for Symbol Grounding.

> *"How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?"*
>
> The Symbol Grounding Problem (Harnad, 1990, p. 1)

4.0. Searle's Chinese Room and the grounding problem.

An early output of the then nascent AI research field was the Physical Symbol System Hypothesis that aimed to outline the foundational principle for studying general intelligence (Newell & Simon, 1976). According to the hypothesis, a physical symbol system consists of physical patterns, known as symbols, which can be combined to form more complex symbol structures. A system containing one or more of these symbol structures, is manipulated by a set of processes that are also instantiated by symbols and structures to generate new symbol expressions. The hypothesis asserts that a suitable physical symbol system in this mould is both necessary and sufficient for the existence of general intelligence (Newell & Simon, 1976, p. 116).

> *"By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By "general intelligent action" we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behaviour appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity."* (Newell & Simon, 1976, p. 116).

A conclusion to be drawn from the physical symbol system is that if intelligence requires such a system (necessity), the human mind, being intelligent, must be a physical symbol system. Conversely, if such a system is sufficient in explaining intelligence (sufficiency), then replicating the mind would be possible by constructing a physical symbol system that performs the same functions as the human mind.

Anti-computationlaist like John Searle found this assertion objectionable and they challenged the idea that intelligence could arise solely from physical symbol processing. For the anti-computationlaist, the glaring difference between the mind and a physical symbol system was that the mind appears to process semantics while the proposed physical symbol systems are syntactic engines. The prevailing thought among computationlaist was that if you took care of the syntax, the semantics would take care of itself (Haugeland, 1985, p. 106). Searle set out to show that a physical symbol system was not sufficient in accounting for semantics. He aimed to show that this would disprove the physical symbols hypothesis by negating its sufficiency claim. To this end, he devised a thought experiment known as the Chinese Room Argument (Searle, 1980).

Searle imagined himself locked in a room with batches of Chinese characters. Although he does not understand Chinese, he is given a set of English rules that allow him to manipulate the characters from different input batches to generate responses that are indistinguishable from those of a native Chinese speaker. In this analogy, Searle in the room plays the role of a physical symbol system. Despite becoming proficient at following the rules to produce the right Chinese output, Searle argues that he still does not understand Chinese. Searle aims to demonstrate that merely following rules for manipulating symbols is not the same as understanding. For Searle, this thought experiment shows that symbol manipulation is not sufficient to account for understanding.

Searle argues that this is analogous to how computers function. Computers manipulate symbols according to rules (algorithms) but do not understand the meaning (semantics) of those symbols. Even if a computational system can generate output that appears intelligent, it doesn’t mean that the machine has a mind or genuine understanding.

The Chinese Room thought experiment is a good demonstrator for surfacing the grounding problem in philosophy of mind and artificial intelligence. It follows from the thought experiment to enquire as to how computational symbols can have inherent meaning, rather than just being manipulatable syntax with only prescribed meaning.

The grounding problem refers to the challenge of explaining how symbols or representations used by a system are connected to or grounded in real-world meaning. In other words, how does a computational system, which processes inputs as abstract syntax, link the syntax to actual objects, events, or concepts in the real world? For symbols to be meaningful, they must be linked to something other than just other symbols. A symbol has no inherent meaning unless it is grounded in some real-world experience. Without grounding, systems are merely manipulating syntax without understanding their meaning (semantics). The grounding problem suggests that, without grounding, computational systems, no matter how sophisticated, remain incapable of understanding.

Searle believes that the interesting thing about the mind is how it solves the grounding problem. If the mind is doing computations, it is a trivial feature of the mind. According to Searle, computation is observer imputed. Almost any system could be said to be interpreted as performing computation with the right level of insight and skill by an observer. The molecular movements in a wall could be interpreted as computation (Searle, 1990). If the mind is doing computation, the interesting part of the mind is the processes by which it self-attaches meaning.

In this case what is needed is an account of the mechanics by which mental syntax is causally related to and imbued with meaning with semantic properties that are grounded in the world. Uncovering the mechanics for how mental syntax gains semantic properties in the mind is critical to the AGI agenda because it can provide a path around the symbol grounding problem.

The chapter will explain the mechanics of mental semantic grounding by introducing the Analog Principle as a causal theory of mental content. It will argue that a complete computational model of the mind should include the Analog Principle to ground computational items.

Section 4.1 will revisit key points from previous chapters and emphasize that computational self-sufficiency is essential for the form of general intelligence valued in the Grand Ambition. Computational self-sufficiency implies a strict prohibition on external agents in problem-solving. This poses a significant challenge in solving the grounding problem for computational models of general intelligence. Section 4.2 will introduce a natural account of how the mind acquires semantic properties, based on a causal theory of mental content. This approach suggests that qualitative states function similarly to analog models in engineering in grounding meaning into computational items. Section 4.3 will define this idea as the Analog Principle and evaluate it against common critiques of causal theories of mental content. Finally, Section 4.4 will apply the Analog Principle to address issues raised by Searle's Chinese Room argument and demonstrate how this principle can resolve the shortcomings highlighted in the thought experiment.

4.1. The Grand Ambition and the grounding problem.

4.1.1. The sufficient properties of the Grand Ambition to the grounding problem.

A few points worth bearing in mind were highlighted in the previous chapter. They are worth recapping because they are an important introduction for the discussion in this chapter.

The AGI researcher labouring under the Grand Ambition brief seeks general intelligence that is functionally comparable to human intelligence. The key word in the demand is 'functionally.' AGI research seeks to realize intelligence that functions comparably to human (cognitive) intelligence.

The preceding chapter identified three jointly sufficient properties for the type of general intelligence that is sought in the Grand Ambition. To sufficiently qualify as general intelligence, competence in solving familiar and new problems, must be supported by three further demands.[29]

- **Access to problem defining inputs:** Processed from relevant input i.e., percepts that define the problems that must be solved.
- **Appropriate function:** Processed by way of an appropriate problem-solving process i.e., a process that is reliable, adaptable to fit context, and that can be credited with success for generating the correct output.
- **Self-sufficiency**: Processed with an autonomy that can be credited with instantiating the appropriate solution.

The mind satisfies the three jointly sufficient requirements. If states in the mind, serve to realize conditions that are both necessary and minimally sufficient for general intelligence, then replicating the functional roles of mental states is important for AGI. This is the sense in which the AGI agenda should seek functional comparability with human intelligence.

Yet, a large section of AGI research is proceeding without any commitment to replicating the functional roles in the mind. They might point to the success of output-centric models in realizing artefacts with artificial narrow intelligence. ANIs are not purposefully engineered with regard to recreating mental functional roles. They are output focused, yet they are very successful in reproducing domain restricted intelligence. They might also sight examples of theoretical models such as the Quantum Turing Machine and the Hutter's AIXI, or projected frontier models like LLMs as examples of output-centric models that have been touted to work in principle as AGI without any attempts at homology with the functioning of the mind states.[31]

The previous chapter weighed some of these output-centric models of general intelligence, against the sufficiency checklist to determine if they will qualify as instances of general intelligence in the Grand Ambition mould.[32] They all failed. Specifically, they failed the third requirement on the checklist.

[29] Refer to chapter 2.

[31] Refer to chapter 3.

[32] Refer to chapter 4.

The third sufficient property is a demand for the type of autonomy than can be credited with owning problem-solving. The demand implies a strict prohibition on extrinsic control, or ownership for the computational functions that support general intelligence. The output-centric models (such as the QTM-AGI, AIXI, projected LLMs) fail to meet this dictate for computational self-sufficiency. They all rely on extrinsically formulated or finetuned instructions for computational processing to realize the optimum output. In all these models, the represented computable items in whichever form they appear in the engineers' processes of choice have prescribed meaning. The artefacts' problem-solving competency is executed through algorithms designed by extrinsic agents to manipulate symbols with prescribed meaning. None of these artefacts are computationally self-sufficient.

This handicap in meeting the negative dictate is common to all approaches to general intelligence that focus on input -output mapping. In being agnostic about the processes that realize human-like output, the output-centric engineers must ensure that whatever processes they chose to model intelligence have been carefully designed or calibrated to produce output that matches human intelligent output. This includes ensuring that the computational items of their choice are assigned meaning that cohere with the phenomena in the world that is to be computed, and that the manipulation of these items is designed to realize an output a human would deem rational.

The mind is not handicapped in meeting the negative dictate for computational self-sufficiency. The mind is not reliant on extrinsic agents to ground meaning or calibrate coherence with worldly phenomena into mental states. The mind is capable of self-grounding meaning from worldly phenomena into its own mental states.

4.1.2. The Grounding problem.

It is apparent that the mind is doing something akin to computation. Mental states have content about the world that is presented in some intrinsically accessible form, and this mental content is manipulated through processes that are underpinned by a neurophysiological substrate that appears to be doing syntactic processing. This view of the mind is in keeping with the description of computation as the representation of worldly information and the manipulation of these forms under syntactic rules that transform the information into a desired form. This analogy is sustained whether one takes a connectionist, classic computational, or some hybrid view of the underlying structure of the mind. The mind does appear to be doing computation.

Computation is the manipulation of symbols by way of syntactic rules. Both symbols and the syntactic rules retain conformity with affairs in the world that are to be computed. Symbols have semantic properties that link represent the world as being a certain way. Syntactic rules manipulate the symbols on the basis of their form and not their semantic properties. Despite this, the manipulation of the symbols coheres with the semantic structure in the world that the symbols represent.

If all computation is the rule-based manipulation of formal syntactic forms that is oblivious to semantic form, how does computation track and match the rationality of the phenomenon in the world that is being computed? How is worldly meaning imputed onto the syntax that informs mental processes in the mind?

It has been highlighted that artefactual computation circumvent the grounding problem by relying on extrinsic agency. Built computation owes its internal coherence to the world to purposeful designers who impute meaning onto the system. The symbols in artefactual computation don't have inherent semantic properties. If these semantic properties exist anywhere, they are in the mind of the programmers and the users. The programmers of computers purposefully ensure symbols and their manipulations will closely cohere with the natural system that they intend to model.

One would be hard pressed to find a computational cognitivist that would suggest that mind has the same extrinsic purposeful programming as an artefactual computer. If the mind is performing computation, there must be an explanation of how meaning related to the world finds expression in the mind's underlying syntactic substrate. The artefactual systems rely on external agency to link syntax to semantics. If the mind is a syntactic engine, how does the mind's syntactic manipulation cohere and remain true to events in the world?

4.1.3. Fitting semantics in mental computations by way of causal implication.

Up until this point, folk psychology had appeared to be a powerful argument for the mind as a computer. Now it raises earnest questions for the computational cognitivist. The computational cognitivist leaned heavily on folk psychology to argue that the mind was like a computer. By way of first-person experience, it is apparent that thoughts and ideas are about things in the world. If the computational cognitivist is correct, the question to resolve is where does the mind acquire the semantic properties that define mental processes?

The intuitive answer is that the mind's syntactic elements acquire semantic properties through a causal relationship, rather than by representational similarity. The shape of a symbols in computation is determined solely by its fit in the syntactic structure, not by its semantic properties. Thus, a symbol (a semantic bearing syntactic structure) does not derive its meaning from resembling the thing it represents. The fact that the symbol shape is defined by conformity to syntax structure and not by resemblance to worldly items suggests a causal theory of mental content as an explanation of how syntax are imbued with meaning.

Causal theories hold that mental content gets its meaning through a causal relationship between the worldly item and its mental representation (Stampe, 1977); (Fodor, 1987); (Dretske, 1981). The item in the world causes the symbolic representation in some way. By the account of causal theories, semantic properties are causally imparted on the mental syntactic operations by worldly items and dynamics. It is through this causal relationship between the worldly phenomena and the syntactic operations that the syntactic structure gains the meaning that defines it as a computational system.

The problem causal theories of mental content must contend with is in giving a satisfactory account of the causal relationship between worldly phenomena and the underlying syntactic structure in the mind. The most obvious solution would be to argue for a symmetric relationship between the worldly phenomena and the syntactic structure that bears the mental content. This is improbable. The occurrence of false beliefs implies that certain mental content can be caused by worldly items or dynamics other than what the mental content represents. For example, it is not unreasonable to suggest that one can see a sheep and mistakenly form mental content that represent it as a goat. Or one can hold a mental representation of a goat caused by a hallucinogenic state or a mental disorder. In none of these cases is the semantic property of a goats caused by a symmetric worldly experience with a goat.

There is a second problem that causal theories of mental content must contend with. Folk psychology appears to indicate that mental activity is in some way propelled by semantics. It is the meaning of concepts that appear to drive the direction of computation. The mind appears to be more of semantic engine than a syntactic engine. Causal theories of mental content must not only explain how semantic properties find their way into syntax, but also how they appear to be causally efficacious in symbol manipulation.

Jerry Fodor's response to accounting for how semantic properties causally relate to symbol manipulation has been to argue that formal activity implements intentional mental activity through non-basic causal laws (Fodor, 1990, pp. 137-159). In his explanation, Fodor makes a distinction between strict and non-basic causal laws. Strict causal laws are such as are found in physics (and only physics). They imbue a nomologically sufficient condition for the satisfaction of a consequent event (p.144).The non-basic causal law, on the other hand, are such that are found in all the other special sciences, such as psychology (p.144). The non-basic causal laws are non-syntactic, they state what will happen given that all factors remain equal. As an example, imagine a **law X** to mean that **'snow forms on elevations of 5000 meters above sea level'**. For the sake of this example, we can imagine that anytime you have an elevation of 5000m above sea level, you will have a nomologically sufficient condition for snow fall. Law X is a strict causal law. Now, imagine **law Y** states that **'mountains form snow.'** Law Y is a non-basic causal law such as found in special sciences, in this case geography. Law Y only has causal efficacy if the factors in law X are met. Not all mountains rise above 5000 meters above sea level. Being a mountain is not nomologically sufficient for snow. The factors that the causal efficacy of non-basic laws rely on are a mediating mechanism. Non-basic laws require mediation by intervening mechanisms operating under the strict (syntactic) physical laws. Law Y can only mandate snow if the conditions that satisfy Law X are met. Law X is a mediating mechanism. According to Fodor, worldly meaning is imparted on syntax as non-basic causal laws. Semantic properties are causally efficacious only if they are implemented in an appropriate physical system in an appropriate state.

Fodor intends for his description of non-basic causal laws to be a psychological account of how semantic properties fit in the mind. However, it seems to work just as well as an appropriate and satisfying explanation for fitting semantic properties into the causal chains of artefactual computing too. The semantics in the programmer's mind are implemented in a physical system as non-basic causal laws. Programming computers is the act of generating conditional laws that are to be executed by a formal physical system given the system is in an appropriate physical state. E.g., An instruction like *'If Mountain > 5000ft, then snow.'* could be an instruction inserted in a computer program. The line instructs the computer to impute a certain property if the right conditions are met. In this case the conditional mandate to project the occurrence of snow if the mountain is higher than 5000 feet. The programmer's computer program is a set of conditional operational laws such as these, that should be executed as syntactic operations given that the right conditions are met, and the system is in an appropriate state. They impart meaning on the syntax by eliminating undesired interpretations in exerting conditional causal efficacy.

Fodor's explanation manages to get folk psychology back on the side of the CTM. It conforms to our folk psychology intuition by putting semantics in the driver's seat of computation. If intentional states embody non-basic causal laws that have conditional efficacy over syntactic operations, it would explain why it would appear that mental content in our mind has causal efficacy. It is a satisfying account of how semantics might be of causal relevance in computational systems.

4.1.4. A glaring gap in semantic – syntax accounting.

While Jerry Fodor has managed to explain how semantics can have causal efficacy in syntactic operations, what his explanation lacks is an explanation of the mechanism by which semantic properties are introduced into mental syntactic systems as conditional causal laws. It might be accepted that Fodor is right in stating that the intentional mental states have conditional causal efficacy that is implemented if the physical conditions are right (Fodor, 1990). The question is, how do semantic properties gain this conditional causal efficacy in the mind?

We know that in an artefactual system it is done through purposeful design. The programmer purposefully inserts semantic properties into a system by crafting a set of conditional laws that define meaning in syntactic structures to ensure there is coherence between semantics from the world, the syntax, and the programmers' intentions.

By what mechanism are semantic properties introduced into the mind in a way in which they have conditional causal efficacy over syntax of the mind? Surely it cannot be by purposeful design. Without a mechanical explanation of the process, Fodor has only managed to push the problem back a step. The account needs an explanation of the mechanics of semantic - syntax coupling in the mind to be complete.

Vincent Müller highlights three important levels of description when explaining a computer; (a) the physical level or an account of the actual realization of the computer (b) the syntactic level of the computational algorithm/s, and (c) the symbolic level, or the content being computed (Müller, 2024).

In artefactual computing, we have answers for all three: we can see what the syntax is doing (b), we know the semantics are imputed from the minds of the designer and users (c), and we can provide a physical explanation for how syntax and semantics are linked (a). A programmer, acting as an external agent, connects semantics to syntax by designing conditional operational laws that ensure the system executes syntactic operations under the right conditions.

However, when it comes to the mind, we don't have all these answers. Folk psychology gives us an account of the semantic level (c), and neuro-cognitive science provides some understanding of the syntactic level (b). What's missing is (a); a physical explanation of how semantics and syntax are coupled in the mind.

Without a clear mechanical explanation for how semantics and syntax are linked in the mind, the claim that the mind performs computations can be questioned. This is exactly the point of Searle's Chinese Room argument. For thinkers like Searle, the role of an external agent in grounding artefactual systems represents such a fundamental difference from the mind's self-sufficiency that the analogy between mind and computer becomes untenable. This stark difference is reason enough to abandon the comparison between the mind and a computer.

However, anti-computational cognitivists set too high a standard when they suggest we should abandon the computational metaphor of the mind. Computation aligns well with our common-sense understanding of the mind. While Fodor's explanation for how the intentional mind interacts with mental syntax is incomplete without a natural explanation of the underlying mechanisms, this gap shouldn't invalidate the parts of the theory that work. In calling for the complete rejection of the computational metaphor, anti-computationalists risk discarding valuable insights. What's needed is an account of the mechanism by which syntactic processes are causally affected, and imbued with meaning from, semantic properties that represent the world.

The next section will describe the mechanics of mental symbol grounding. It will propose a natural account of a causal theory of mental content.

4.2. The mechanics of computational grounding.

Causal theories of mental content suggest that the meaning of mental content arises from a causal relationship between an external object and its mental representation (Stampe, 1977); (Fodor, 1987); (Dretske, 1981). In this framework, the worldly object causes the mental content. Or in other words, the semantic properties of a mental representation are causally imparted by real-world items and dynamics. This causal relationship allows the syntactic structures in the mind to acquire the meaning that defines them as a computational system.

A key challenge for causal theories is explaining how a world that appears continuous can impart meaning on a computational structure expressed on discrete units. The conflict is that the world is experienced as continuous, yet computation relies on discrete measurements. A measurement is a quantized approximation of a continuous phenomenon. For instance, heat is a continuous signal, but to compute its properties, it must be quantized into discrete units. In the Celsius scale, a convention adopted to support communicating and computing aspects of heat, a degree Celsius is a quantized representation of one-hundredth of the temperature change between the freezing and boiling points of water. However, measurements like the Celsius scale are arbitrary constructs with meaning derived from consensus. There is no inherent link between the representational unit 1°C and heat itself. The meaning of this unit arises from its conventional use by users who agree that 1°C represents some structural value of heat.

Questions can also be raised about the accuracy of breaking down continuous phenomena into discrete units. How well can a discrete item represent a continuous signal in the world, especially when that signal exists alongside other signals in a noisy continuous and variable whole? This challenge of dividing the continuous physical world into discrete computational units is one reason Deutsch advocates for the use of quantum processes in proving the Church-Turing-Deutsch principle (Deutsch, 1985). Quantum processes offer a more physically accurate representation of complexity and knowledge compared to classical computing, which relies on manipulating discrete units governed by physical laws.[39]

These challenges suggest that a third party is needed to establish how the continuous world is divided up into discrete units. While this external role is not an issue for artefactual computation, which unabashedly and explicitly involve an external designer, it raises questions about how causal theories account for meaning in mental processes without the role of an extrinsic agent.

Causal theories of mental content suggest that the mind manages to navigate these problems without recourse to a third party. At the most fundamental syntactic level, the mind can be reduced to binary neural signals' 'on' and 'off' firings. The computational symbols in the mind are tokenized expressions or patterns that arise from this neural activity. Insights from artificial neural networks suggest that these tokens are most likely built on distributed weights and biases in the interconnected neurons' firings (McClelland & Ralph, 2015), (Edelman, 2002). While the exact details are not crucial here, it is evident that at its core, the syntactic level of the mind is composed of discrete physical operations.

In framing the problem, we can say that the world is experienced as continuous, while computers process information in discrete units. Without an external designer, causal theories must explain how continuous information from the world is processed by the mind in a self-sufficient way to ground meaning onto mental syntax.

4.2.1. The role of qualitative states in mental grounding

Here we have two levels of explaining the mind that require integration. Folk psychology provides a satisfying, meaning-driven explanation of how the mind operates. In this explanation of the mind, underived meaning from the world is grounded into mental activity in causal process that involve continuous qualitative states. The primary example of this is in veridical perceptual experiences and how they furnish knowledge of the world that informs mental activity. On the other hand, neuro-cognitive science offers a syntactic explanation that describes the mind as a physical system that fundamentally operates through binary or possibly quantum supported physical neural activity.

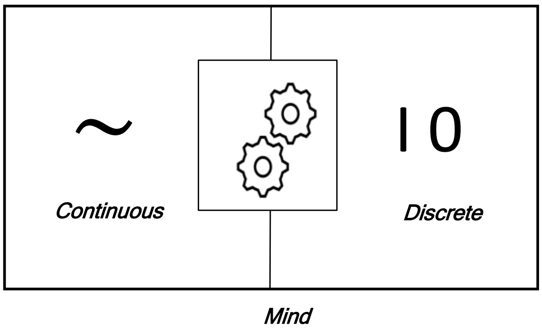
What is lacking is a clear physical account of how these two explanations converge.

[39] It should be noted though, that while quantum computing might enhance information capture and processing efficiency, it would still rely on external grounding to assign meaning to its underlying qubit activity.

By combining the folk psychology view of mental content with the neuro-cognitive understanding of the brain's structure, a perspective of the mind emerges where the mind consists of qualitative states that are continuous, and through principles we aim to uncover, anchor meaning onto the underlying syntactic structure of the mind. If this view holds, it suggests the mind has a continuous "front end" and a discrete, operational "back end."

In very simple terms, we can use the schematic model below as an abstraction to illustrate the combined explanations.

Image 4.1. The grounding in the mind.



In image 4.1, the mind is depicted as having both continuous and discrete properties. **The sine wave represents the continuous properties**, which are the qualitative states we experience in the mind. For example, the qualitative states of perceptual experiences. **The binary 1s and 0s, typically associated with digital computing, represent the discrete activations in the neural structure of the mind**. The gears symbolize the mechanical connection between the continuous signals and the discrete elements that we aim to lay bare.

Common sense psychology suggests that in veridical perceptual experiences, the perceptual qualitative states in the mind meaningfully reflect the world and its dynamics. These perceptual qualitative states are in an analogous relationship with the world. This means that they reflect in some meaningful way the dynamics in the world. In this case, "analogous" should be understood in its *ad litteram* meaning as having a similar form but being different in ontic nature.

We can enhance our schematic model to represent the analogism in veridical perceptual experiences.

Image 4.2. Model of veridical perceptual experience in the mind.

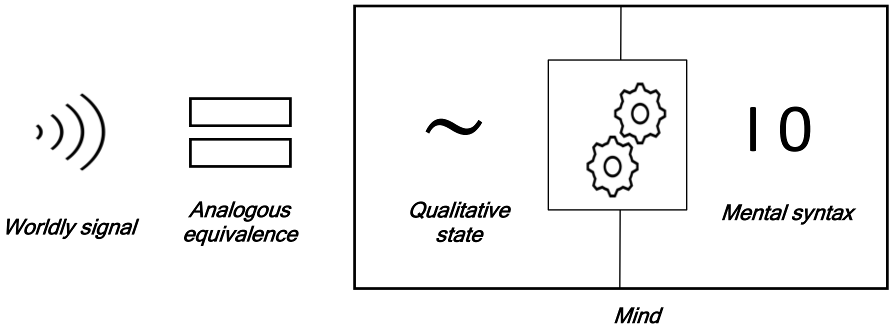


Image 4.2 presents an updated model of the mind to reflect a veridical perceptual experience. The 'worldly signal' represents perceptual stimuli from the external world. It is symbolized by a broadcast icon. These stimuli are in an analogous relationship, that is depicted by the equals icon, with the continuous properties of the mind. The continuous properties are the perceptual qualitative states, represented by a sine wave icon in the schematic. These qualitative states are connected to the syntactic elements of the mind, which are the discrete components of its physical functioning. The syntactic elements are represented by the binary 1s and 0s icon.

At this stage, a causal chain of how meaning is grounded is beginning to emerge. Qualitative states model the world as it appears and, through internal connections, embed this information into the syntactic substrate of the mind.

4.2.2. The importance of qualitative properties.

The emerging causal chain implies an important role for qualitative properties in qualitative states. The emerging causal chain of grounding suggests that qualitative states model the world as it appears and, through physical linkages, embed meaning into syntax. In this view, it is crucial for qualitative states to accurately represent the world.

Qualitative states come in many forms. For example, perceptual experiences include the phenomenal experiences of seeing, tasting, hearing, smelling, touching etc. These experiences differ by type (e.g., seeing is distinct from hearing) but also vary within each type (e.g., different kinds of visual experiences). Despite this diversity, all these experiences share the common feature of being qualitative in the sense that there is something it is like to have the experience.

Although qualitative states are rich in variety and type, they share common properties. They are discernible (able to be perceived), differentiable (able to be distinguished from one another), continuous, and potentially variable. These properties are essential in maintaining the analogous relationship between the representational qualitative state and the perceived phenomena in the world. Discernibility ensures that worldly signals being modelled can be individuated. Differentiability allows perceived signals to be distinguished from other signals. Continuity mirrors the continuous nature of real-world phenomena, allowing for more accurate modelling of worldly signals. Variability captures the dynamic qualities of real-world phenomena, ensuring a more nuanced representation of the world. Together, these properties enable a more faithful and complex representation of the world.

4.2.3. The Analog Principle.

The process of modelling a continuous signal to ground information onto discrete units is a type of analogical modelling. Analogical modelling uses an analog model to indicate or describe a special relationship between an original (target) signal by creating analogies with its features to some other secondary signal in a totally different medium (Gentner, 1989). In engineering, the most frequent use of analog models is to represent or convert a continuous worldly signal to a variable physical quantity. For example, an analog recording is a recording that models continuous sound waves with electrical signals to convert or store the recording as discrete data.

It's important to clarify that this usage of 'analog' differs from its traditional use in the philosophy of mind, particularly in Computational Theory of Mind (CTM) and Representational Theory of Mind (RTM). In these contexts, 'analog' typically refers to the format in which mental content is represented. Analog theories of representation hold that mental content has spatial representational properties. It states that mental content is presented in some form that is akin to pictures.

The word 'analog' here should be read in the engineering sense. An analog model in this reading refers to the analogous representation of a continuous worldly signal for the purpose of transferring meaning from the structural features of the modelled signal onto a different medium. This includes mapping meaning from modelled signals onto computational items. Computational items are basic forms or values used in algorithmic processes to understand or solve problems. These forms can be measurements or other types of physical discrete forms that can be used in computation.

For example, a clock uses the continuous angular movement of its hands as an analog model, of the movement of shadows on a sundial.[42] This analog model transfers meaning relating to the continuous natural phenomenon onto computational items, in this case, units of time. Other

[42] The movement of shadows on a sundial are an analog model themselves. They are an analog of the apparent movement of the sun *vis-à-vis* the earth's rotation around its axis.

examples include a mercury thermometer, which uses the expansion and contraction of mercury to model temperature changes, and a compass, where the needle's movement models the earth's magnetic field. In all these examples, the analog model represents a worldly continuous signal to transfer meaning onto discrete items that can be used for further computation.

Again, the key distinction between the use of ‘analog’ here and its use in the philosophy of mind is that, in engineering, it refers to the functional role of the analog rather than the format of representation. The form of the analog model, whether that be the angular movement, thermodynamic expansion, or something else, is not important. What defines an analog in engineering is its function in transferring meaning from the model onto to discrete computational items.

However, while the analog in this context is agnostic to form, an analog model is still reliant on qualitative properties. To accurately model a continuous worldly signal, the analog model must maintain an analogous relation to the signal. To this end, the properties of discernability, differentiability, continuousness and variability are important. To model a worldly signal accurately, the signal must be discernible (able to be perceived) and differentiable (able to be distinguished from other signals). The analog model must also match the continuity and variability of the parent signal. For example, the movement of the hands of a clock match the discernible, differentiable, continuous and variable movement of the sun across the sky in some meaningful way. The qualitative properties are essential for capturing the important structural features of the modelled worldly signal and transferring that meaning onto computational items.

The process of using an analog model to transfer meaning relating to the structural features of a modelled worldly signal onto computational items is what this thesis terms the Analog Principle. The systematic execution of this principle in physical systems is referred to as the Analog Mechanism. A clock, thermometer, and compass are all examples of analog mechanisms operating under the Analog Principle.

4.2.4. The Analog Principle and computing.

The Analog Principle is a core principle in computation, though its significance has gone unarticulated, and of recent, underappreciated and overlooked. Early computers relied overtly on analog mechanisms to transfer meaning relating to worldly signals onto computational items that could be processed by the system's physical syntactic and mechanical structures. To automate computing, these systems required physical models that represented the phenomena providing the data points to be computed. It is because they relied overtly on these physical models, that they became known as analog computers.

Modern digital computers also depend in an indirect way on the Analog Principle. However, this reliance has been obscured by advancements aimed at improving efficiency. The thesis will expound on this in more detail later. In the interest of narrative clarity, it is crucial to begin by discussing the more explicit use of analog models in early computational systems. A good place to anchor this discussion is on Lord Kelvin's Harmonic Tide Analyzer.

Lord Kelvin's Harmonic Tide Analyzer.

Early dependants on the seas would have noticed that tides behaved in complex patterns. The patterns in which the shores rose and receded was a matter of interest to more than intellectual curiosity. Understanding the tides was important for livelihoods; it informed, fishing, farming, commerce, travel, and more.

Sir Isaac Newton offered an early insight into the forces that were at play in tidal movements in his *Philosophiae Naturalis Principia Mathematica* (Newton, 1687). In it, Newton states that ocean forces are governed by gravity from the sun and the moon. Newton identified a primal force in the complex patterns of tidal movements, but it was not the full story. There were other auxiliary factors at play in accounting for tidal dynamics.

In 1775, Pierre-Simon Laplace, a French astronomer and mathematician, built on Newton's contribution by developing a dynamic theory of tides. Laplace's theory took into consideration the earths tilt, the rotation of the moon, the revolution of the earth, The wobble of the Earth's axis, the natural oscillation periods of the ocean basin, the natural oscillation periods of coastal bays and other factors (Cartwright, 2000, pp. 73-75) Laplace's theory was a step closer to a full picture of the variables that informed tidal patterns. By 1876, scientists had identified at least thirty-seven variable factors that controlled the tides for most locations. By calculating measures of each of these factors at a given time, scientists could remarkably close to accuracy predict tidal times and heights. (Cartwright, 2000, p. 88).

However, such calculations were laborious. The factors were constantly shifting and therefore required that the calculations had to be done several times to get a real time snapshot of the tidal movements. The values of the factors were also location dependent, which meant that different calculations had to be done for separate locations. What was needed was some way to automate the computations.

This is the problem that Sir William Thomson, later to be known as Lord Kelvin, set out to solve. He set out to build a mechanical machine that could compute a prediction for the tides in various locations. Lord Kelvin identified ten variable factors that he thought were essential for a workable prediction of the tides in most places. The primary challenge Lord Kelvin had to overcome was to devise how his machine could represent the ten dynamic factors in the world as computable items. In other words, how could he make the internal movements of his mechanical contraption have meaning that cohered with the dynamics in the real world that he had to compute to project their transformation?

Lord Kelvin solved the problem through a contraption that he called a Harmonic Tide Analyzer. The Harmonic Analyzer had ten wheels that could be configured to match in movement the dynamism of each of the ten constituent factors that Lord Kelvin identified as important for predicting the tides. When a handheld crank was turned, the wheels put on a performance that was faithful to the real-world forces they were modelling. Each wheel had a shaft attached to it that transferred the motion of its wheel to a gear-train that was arranged to combine all the elements in a manner that was faithful to the interaction in the real world of the real forces being modelled. The gear-train fed into an output gear that traced a tidal curve onto a paper. The traced tidal curve on the paper represented the desired output. If the right adjustments for each variable factor in a given area was configured into each of the wheels, the Harmonic Analyzer was able to draw the tidal curves of that given area for one year in less than four hours (Cartwright, 2000, pp. 97-100).

The machine was a success. So much so that similar machines functioning under the same technology were still in use well into the 20th century. In 1942, Arthur Dodson overhauled two machines inspired by the Harmonic Tide Analyzer to include twenty-six tidal factors. He used the two machines to predict the tides for the Normandy invasion in World War II (Parker, 2011).

The Harmonic Tide Analyzer is an early example of an analog computer. An analog computer uses a model of the discernible, differentiable, continuous and variable aspects of a worldly phenomenon to transfer meaning onto computational items. The ten wheels in motion in Lord Kelvin's machine constituted a mechanical model of the tidal dynamics in an area of interest. In this sense, the model is an analog of the real-world problem to be computed. The modelling of the real-world problem serves a purpose. In being a faithful recreation of the real-world phenomena in some simplified, but meaningful sense, the model has meaning which is like the real-world phenomena. This meaning can thus be transferred onto the inner workings of the computational process.

The Harmonic Analyzer transferred meaning onto its computational processes through the shafts that run from the wheels to the gear-train. The gear-train was arranged in such a manner as to constitute a computational function that translated all the input from the shafts into the motion of a single output gear. The shafts thereby conveyed meaning into the Harmonic Analyzer's computations by ensuring that the computations were about a real-world phenomenon. In this case, the tidal dynamics of a given area which was modelled by the analog model.

Lord Kelvin's Harmonic Tide Analyzer utilized the analog principle to ground its computational items. The analog model in the Harmonic Tide Analyzer explains how meaning that relates to an area's tidal dynamics is introduced into the Harmonic Analyzer's computational structure. The model of the tidal dynamics transfers structural meaning relating to tidal dynamics onto the cogs and gears of the gear-train that execute the Harmonic Analyzer's computations.

The analog principle predates the Harmonic Analyzer. All early computers employed analog mechanisms to transfer meaning relating to signals from the world onto computational items that could be manipulated by the mechanical structure of the system. To automate computing, you had to build a physical model of the phenomena that supplied the data points to be computed.

There are several other examples of early analog computers that were built with working models of the physical phenomena they were built to analyse. Post-Lord Kelvin's Harmonic Analyzer, Hannibal Ford built the Baby Ford Integrator in 1917. It used a physical analog model to extract data to compute the differential equations that governed projectiles. For an even earlier example, there is the Antikythera Mechanism, which has been dated to as far back as 87 BC. It was a hand powered orrery that used gear movements as an analog of astronomical movements (Freeth, et al., 2006), (Pinotsis, 2007). The gears converted the modelled information into data that was used to predict eclipses and track the 4-year cycle of the ancient Greek Olympic games (Freeth, et al., 2008), (Iversen, 2018). The analog principle has long been an essential principle for symbol grounding in computing.

A keen reader might point out that modern computing did away with physical analog models. This might raise questions about the essentialness of the analog principle to computing. In the digital era, computing has transitioned from grounding via modelled continuous signals to grounding through the purposeful encoding of discrete information into digital items.

It must be noted though that the shift to digital computing was done in the interest of automating the analog principle in computing, rather than to do away with it. The innovation of digital computing is a hack that was intended to ease the engineering aspect of adjusting physical analog models to suit different computational projects. The word 'hack' as used in this case has transitioned from informal to formal use in computing parlance. It refers to a convenient but often inelegant fix for a problem.

Digital computing as an innovation owes a great deal to the pioneering work of Claude Shannon in communication and computation theory. To understand how digital computing obscured the analog principle in computing rather than replaced it, it is worth exploring, in brief, Shannon's contribution to the evolution of digital computing.

Claude Shannon's hack of the Analog Principle.

All early computers utilized an analog to ground their essential computational items with meaning that referenced the phenomenon they aimed to compute. However, these machines had one handicap. In relying on a built physical model of the phenomenon they were computing, they were limited problem solving tools. They could only be used to compute the single problem for which their analog model had been modelled and built.

To solve the single domain problem of early computation, Vannevar Bush built the Differential Analyzer, an all-purpose analog computer, in 1931. The Differential Analyzer worked like its predecessors. It had gears and wheels that were arranged to embody the forces at work in the natural phenomenon that was being modelled. These gears and wheels would in turn power shafts and cranks that isolated the structural information from the analog model and linked it to a gear-train that had been calibrated to embody a computational function that produced the desired output. The difference between the Differential Analyzer and the earlier analog computers was that the machine was convertible. For each new problem that the machine had to solve, the machine's analog model could be disassembled and reconstructed to model the new physical system that was to be computed.

The deconstruction and reconstruction of the Differential Analyzer's guts was tedious and physically demanding. Engineers sought to make the disassembly and reconstruction of the internal mechanisms of the analog computer easier by automating the deconstruction and reassembly with electric switches and relays that would realign the bowels of the machine to the problem being solved. The new electrical approach reduced the physical toil of assembling the physical analog models for the machine, but it still commanded a great degree of mental application and time dedicated to trial-and-error approaches to building the right analog model.

This is the problem that Claude Shannon set out to solve in A Symbolic Analysis of Relay and Switching Circuits (Shannon, 1938). Shannon discovered that Boolean algebra could be replicated exactly in electrical switches and relays. True and false corresponded with on and off switches of electric currents, and the Boolean logical operators could be replicated as circuits (Sonni & Goodman, 2018). By reducing switches and relays to symbols, Shannon ensured that the mechanical rearranging of computational hardware to model new problems was obsolete. A programmer could use Boolean algebra to model logic gates that introduced meaning into the computation through conditional statements. Less than a decade after Shannon's paper, analog computers had been replaced by digital computers.

Shannon's breakthrough to digital computing took computing further way from the computational self-grounding. This transition away from analog models in computing has served to increase the dependency of computing on extrinsic agency. The old analog computers used an analog model to transfer worldly meaning onto computational items. While the analog models were designed by extrinsic agents, the grounding was effected through the analog principle. In digital computing, meaning is imputed directly onto computational symbols and syntax through conditional statements in a programming language that is underpinned by Boolean rules. The programmer ensures that computational symbols and syntax are encoded with meaning in coherence with the dynamics of the phenomenon being computed. Digital computing is wholly reliant on extrinsic agency for grounding. This complete reliance on extrinsic agency to ground meaning onto computational items doesn't have any direct parallels in mental processes, and it is what has prompted CTM sceptics like Sayre and Searle to argue that the mind is not doing computations (Sayre, 1986), (Searle, 1990).

Digital computing did not devalue the essentialness of the analog principle in computational grounding. It obscured it, by expanding the role of an extrinsic agent to include grounding. In so doing it made artefactual computing less like cognitive computation.

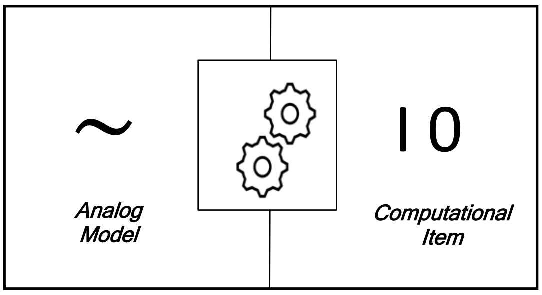
If the analogy of the mind as a computer is to be sustained, it must be demonstrated how the mind effects the analog principle to ground its computational items. To do so it is important to begin by understanding how an analog mechanism works.

4.3. The Mechanics of the Analog Principle.

The mechanics in the analog principle are easy to grasp. An analog mechanism uses an analogous model of a continuous parent signal in the world to encode, ground, or transfer structural information from the modelled signal onto a useful computational form. The transfer is affected through a reliable synching between meaningful structural features in the analog model and the computational item.

One can therefore abstract two parts from the analog mechanism.

Image 4.3. The Analog Mechanism.



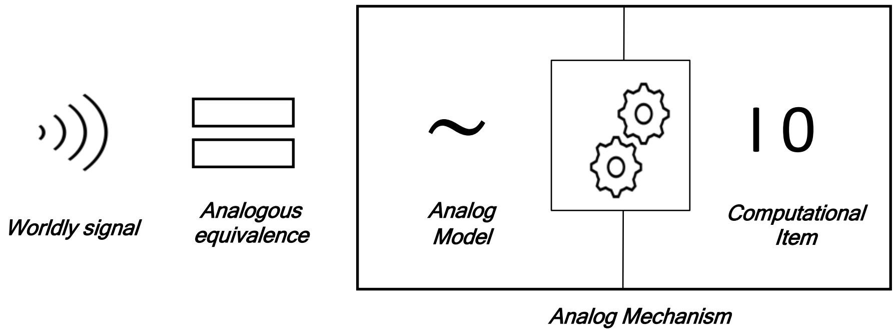
Analog Mechanism

Image 4.3. is an abstraction of an analog mechanism. The mechanism consists of two parts. The first part is the analog model, which is represented with a sine wave. The second part is the computational items. They are syntactic structures that are particular to the computational system. In the image above, the syntax is represented with binary code (1 and 0). The gears represent the transfer of meaning from the analog model to the computational items. In early analog computers, this transfer was done through a physical linking.

It was noted that the analog model is designed to be in an analogous relationship with the worldly phenomenon that is to be computed. That is to say that the analog model should represent in some reliable ways the informative signals from the world.

The schematic model can be further developed to highlight this analogous relationship between analog model and the worldly signals.

Image 4.4. Grounding in the Analog Principle.



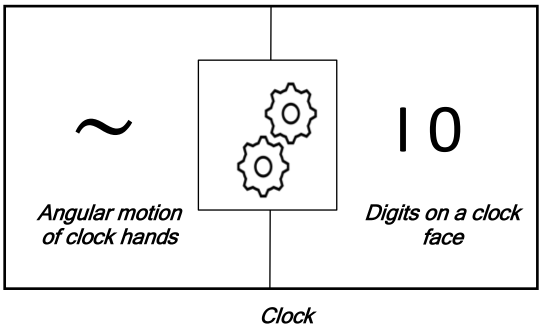
The image 4.4. aims to support the visualisation of the analog principle. **The worldly signal is represented in the image by the Broadcast icon.** Worldly signals are discernible, differentiable, continuous and variable signals from the world that define the phenomena that is to be computed. **The Equal Sign icon represents an analogous relationship**. It indicates that **the analog model represented by a Sine Wave icon** is a simplified but a structurally faithful representation of the worldly signals. The analog model is represented as physically attached to **the syntactic substrate, which is represented by the Binary Digit icons**. This physical connection aims to communicate that the analog model and the computational items are synchronised in some meaningful and reliable way. In describing the Harmonic Tide Analyzer, it was revealed that the syncing connection between the wheels that represented the analog model and the gear-train that represented the computational items was physically effected through shafts. These shafts transferred meaning from the analog model onto the computational items. **The Gears icons in the image represent the transfer of meaning** from the analog model (sine wave icon) to the computational items (binary digit icons).

4.3.1. A clock as an Analog Mechanism

A real-life example of an analog mechanism can help illustrate the two key properties of an analog mechanism and their function. Earlier in the chapter, a clock was highlighted as a simple example of such a mechanism. A clock has both an analog model of a worldly phenomenon and computational items that derive meaning from this model.

The analog model in the case of a clock is represented by the movement of the clock's hands, which model the motion of shadows moving across a sundial. This movement captures essential structural information from the world and converts it into computational items. The computational items are the digits around the clock face. These digits gain meaning from the model. The meaning is important for computing time.

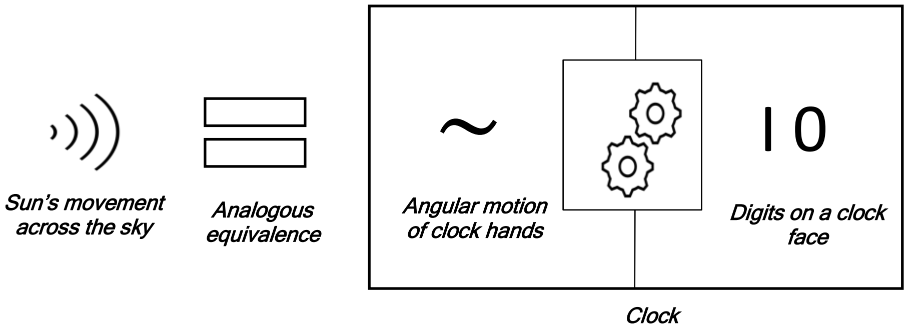
Image 4.5. A clock as an Analog Mechanism.



The **analog model represented with a sine wave** is the angular movement of the clock hands around the clockface. The **computational items represented with binary code (1 and 0) are the digits** 1 through 12 used in modern conventions for calculating time. The gears represent the transfer of meaning from the model onto the digits. This transfer is effected through the clock hand pointing at a digit to embed it with meaning that relates to a time of day.

It's important to remember that the angular motion of a clock's hands on most modern clocks is a model of the shadows moving across a sundial, which in turn models the apparent movement of the sun across the sky. Therefore, the clock's analog model maintains an analogous relationship with a natural phenomenon. In other words, the analog model of the clock meaningfully represents the sun's apparent movement across the sky. We can update our model of the clock as an analog mechanism to represent this analogous relationship.

Image 4.6. The Analog Principle in a clock.



In the clock example above (image 4.6), the worldly phenomenon, the sun's apparent movement across the sky (broadcast icon), is in an analogous relationship (equals icon) with the analog model, the angular motion of the clock hands (sine wave icon). The angular movement of the clock's hands grounds meaning into the computational item (the binary code icon). In the example, the computational items are the digits 1 through 12 located around the clockface. These computational items are used to compute time in the most popular time keeping convention. The digits are encoded with meaning relating to worldly phenomena via the analog model. For example, the unit 12pm has been encoded to mean the period of the day when the sun is directly overhead, by way of the clock's hand pointing at the digit 12 at this time.

4.3.2. The mind as an analog mechanism.

There are clear parallels between the explicated theory of grounding in the mind in section 4.2 and the mechanics of the analog principle explicated in this section 4.3. These parallels can be brought to life by overlaying the schematics used to model the two processes into one model.

Image 4.7. The Analog Thesis of the mind.

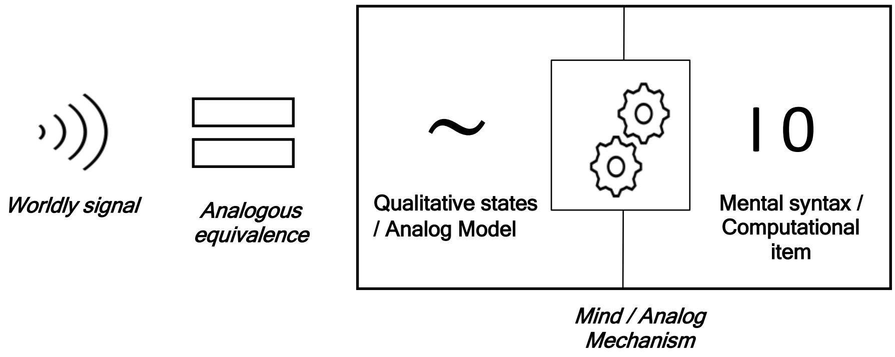

Worldly signal
Analogous equivalence
~
Qualitative states / Analog Model
1 0
Mental syntax / Computational item
Mind / Analog Mechanism

Image 4.7 shows the parrels between the grounding in the mind and in an analog mechanism. The parallels suggest that the mind grounds syntax by way of the Analog Principle. The worldly signals in the image are signals from the world that are discernible, differentiable, continuous and variable. These signals are in an analogous relationship with the qualitative states in the mind. An analogous relationship implies that worldly signals and the qualitative states have comparable qualitative properties, even if it is the case that they are ontically different. That is to say that the qualitative states in the mind represent in some meaningful way the discernability, differentiability, continuity and variability of the worldly phenomena.

The qualitative states in the mind ground mental syntax with meaning. In the case of the mind, the syntax is tokenable items borne on the neurophysiological structures of the mind. Connectionist might view these tokenable items as distributed activations, while classic computationlaist might suggest that it is localized activity. Whichever way one views the tokenable items, what matters is that the tokenized items are a form that is manipulatable by the internal syntactic structure and movement of the system. The tokenable items are encoded with meaning through association with structural information from the analogized representation of the worldly signal.

The Analog Principle Thesis (AP Thesis) is a naturalist causal theory for how the cognitive mind self-grounds meaning that relates to the world onto mental syntax. The AP thesis states that the mind is an analog mechanism. Mental content has semantic properties that relate to the world because it has been grounded by an analog model in the form of qualitative states. For example, the neurophysiological-cognitive items that represent a “tree” in cognitive process have ultimately gained that semantic content through the same principle that the unit ‘12 PM’ gains the semantic content that relates to a specific time of day in which the sun is directly overhead. Both computational items have been imbued with meaning by an analog mechanism that causally associated the computational items with worldly information.

It should be kept in mind that the Analog Principle is agnostic about the form of the analog model. The Analog Principle defines a functional role for the analog model. As we have seen in the various examples, an analog model in an analog mechanism can be the angular motion of a clock hand, the expansion and contraction of mercury in a tube, or the magnetic propelled motion of a needle. The form of the model is not important. What matters is that the analog model matches in some meaningful way the qualitative properties of the worldly phenomena that is being represented. The analog model must match the discernibility, differentiability, continuousness and variableness in a manner that meaningfully represents the worldly phenomena. Therefore, the phenomenal form of qualitative states in the mind are not interesting to the Analog Principle. What matters is that the qualitative properties in the mind represent the worldly phenomena in some meaningful way.

Understanding the mind as an analog mechanism helps resolve the issue of how semantic properties can influence a syntactic process. The challenge arose because computation seems to consist of purely syntactic operations, leaving no obvious place for semantics in the physical system. In artificial systems, this problem was addressed by attributing semantics to the mind of the programmer, an external agent. Originally, the programmer created an analog model to physically transfer meaning to the syntactic structure. After Claude Shannon's digital revolution, this meaning was embedded in the system through logical statements, which function as conditional causal laws. These laws have conditional efficacy that ensures the system retains coherence with the structure and dynamics of the worldly phenomenon being computed. In the mind, the semantic properties are causally self-grounded through the Analog Principle.

4.3.3. Semantic Properties as non-natural, underived meaning.

The Analog Principle thesis of the mind (AP Thesis) is a causal theory for semantic grounding in the mind. It states that the mind introduces semantic properties into its processes by way of an analog mechanism that grounds and calibrates computational items in the mind to have meaning that coheres with events in the world. The AP Thesis is not meant to be understood as a complete account of computation in the mind. It does not describe cognition in its entirety. The AP Thesis is limited to describing how computational items used in syntactic mental computation get their meaning.

The AP thesis is concerned with a very particular type of meaning. First, the meaning must be underived. This means that the content does not gain its meaning from cultural or collective agreement. For example, the red octagonal stop sign used in traffic signage has meaning located in collective agreement. It has what is called derived meaning. The AP Thesis aims to explain mental content with underived meaning.

Secondly, the content must have non-natural meaning. Unlike natural meaning which does not generate falsity, non-natural meaning can be falsifiable (Grice, 1957). An example of natural meaning is how smoke indicates combustion. One can conclude on seeing smoke that it means something is combusting, because smoke entails combustion. Mental content is not natural. It can be falsifiable. Therefore, it has non-natural meaning.

As an example of the falsifiability of mental content, consider reliable misrepresentations. A reliable misrepresentation is a consistent failure in representing an item correctly under certain conditions. In other words, it is a representation that is always misrepresented in the same way. For example, a sheep might always appear as a goat from a certain perspective, or a yellow wall might always appear white under certain light.[43] Reliable misrepresentations would be impossible to account for as content with natural meaning. To highlight this difficulty, consider that a natural meaning account of mental content would proceed as follows:

A syntactic structure "X" means X because "X" is naturally entailed by X.

There does not seem to be room in this accounting for reliable misrepresentations. If "X" means X because X entails "X," then why do some X's cause "Y" under certain conditions; why do some sheep cause syntax that have the meaning of goats consistently under certain conditions? The occurrence of reliable misrepresentation in mental content means that mental content cannot be accounted for as containing natural meaning in causal theories of mental semantics.

Angela Mendelovici not only agrees that mental content cannot be accounted for as containing natural meaning, but she also argues that a causal theory of mental semantics should allow for cases of reliable misrepresentation (Mendelovici, 2013). Mendelovici is responding to arguments that aim to explain away reliable misrepresentations as failures or glitches in causal theories that apply natural meaning to mental content. According to Mendelovici, semantic properties could not possibly reflect natural meaning. As an example, to put her point across she uses the concept of heaviness. A mental item "X" that represents intrinsic heaviness, is a misrepresentation because nothing can be intrinsically heavy. An item that is heavy on earth is not necessarily heavy in other parts of the universe. Nevertheless, it is a useful reliable misrepresentation because the concept of heaviness will be consistently caused by the same things on earth. Reliable misrepresentations appear to be useful features of cognition, rather than glitches in a causal theory of semantic properties that have natural meaning.

The AP Thesis comfortably accommodates cases of reliable misrepresentation. The thesis does not mandate a natural entailment between the worldly signal that originates meaning and the syntactic structure that gains the meaning. The AP Thesis argues that syntactic structure are grounded via a mediating analog model that is in an analogous relationship with a worldly parent signal. The AP thesis of mental content very loosely interpreted takes the form of:

[43] Reliable misrepresentations consistently misrepresent in the same way under certain conditions. In this sense they are different from a hallucination or an illusion. Some illusions are regular and predictable, but they differ from reliable misrepresentation in that they are compatible with the overall veridicality of a well-functioning system (Mendelovici, 2013). For example, the duck-rabbit illusion is supposed to present as either a duck or rabbit depending on what the observer chooses to focus on.

> **Worldly item X is modelled by analog “X” which encodes structural features “”X”” in a syntactic structure.**

Given the imperfections in nature, it is certainly imaginable that there will be conditions in which the fidelity of the analogous relationship between worldly item and analog model might be strained or corrupted along this causal chain. In such conditions, the analog model might consistently misrepresent the parent signal under certain conditions. An example might elucidate the point. Imagine a student recording a lecture on a tape recorder. In this case the voice projected by the lecturer represents the worldly signal, and the voice captured on the tape is the analog model. For the most part, it would be expected that the voice recording is a faithful representation of the lecture. However, it can be imagined that certain factors might cause disruptions on the recording. There might be other noise signals in the environment that might obscure the voice recording. Or a physical fault on the tape that might corrupt the recording. In such cases, the recorded lecture will misrepresent the affected part of the lecture. Anyone who listens to the voice recording will have a misrepresentation of the affected part of the lecture.

Likewise, if the mind is an analog mechanism, then it is expected that certain conditions might affect the fidelity of the analogous relationship between the parent signal and the mind’s analog model. Or just as likely, the disruption might occur between the analog model and the transfer of meaning onto the computational items. In both cases, systematic problems will affect the fidelity of the symbol grounding process. An account of reliable misrepresentation in the AP thesis might proceed as follows:

> **Worldly item X is modelled with a loss in fidelity as “Ɣ” which encodes structural features with a loss in fidelity as “”Y”” in a syntactic structure.**

In this account we can see how imperfections in the causal chain of the AP thesis’ processes can cause a reliable misrepresentation.

The AP Thesis can also account for reliable misrepresentations as useful design feature in a causal theory of semantics. Mendelovici argued by way of an example that reliable misrepresentation should not be explained away as failures of a causal theory of semantics with natural meaning. She used the example of heaviness to support her point. A mental item that represents intrinsic heaviness cannot have natural meaning because nothing can be intrinsically heavy. Mental content that has the meaning of heaviness appears to be a reliable misrepresentation that does not appear to have occurred in error because it is useful for cognition.

The AP thesis states that the mind grounds its computational items through a mediating analog model of the natural world. This analog model retains a fidelity to the relational structures that exist in the world, and it encodes them onto the computational items. For example, the concept of heaviness is defined by its structural relationships. Heaviness is a relative property brought to life by other properties such as gravity, strength, location etc. The qualitative experience of heaviness caused by lifting a heavy book on earth is part of an analog model that mirrors all the relational properties that inform the real-world event. This experience encodes the property of heaviness into the syntax that are used to define or make computations related to the book. For example, the agent can use this information to compute how far they can throw the book or how high they can lift it. If the structural relationships in the world are changed, as would be the case if the same book were experienced on a dwarf planet, the new state of the world would be modelled by an analog model informed by the new structural relationships. These new relational structures might combine to cause a qualitative experience of lightness. The new qualitative experience will recalibrate the syntax used to define or make computations related to the book with the property of lightness. The agent might now compute that they can throw the book further or lift it higher. The analog model, in this case the qualitative experience of heaviness or lightness, ensures that computational syntax is grounded and coheres with how things are in the world. The AP thesis is a satisfying explanation for how mental content gains non-natural meaning.

The AP Thesis is a natural explanation of the mechanics of how mental syntax gain non-natural, and underived meaning.

4.4. The Analog Principle and Searle’s Chinese Room.

Searle's Chinese Room argument challenges the idea that the mind works like a computer (Searle, 1980). Searle argues that computation alone cannot explain how the mind handles meaning (semantics). In his thought experiment, the "Chinese Room" represents a computer which processes Chinese characters to produce correct output without understanding their meaning. Searle aims to shows that, despite following algorithmic rules, the person inside the room does not actually understand Chinese. This he states, demonstrates that computation is not enough to account for understanding or semantics in the mind.

Searle’s Chinese Room argument is dependent on a definition of computing as strictly syntax, and no semantics. His starting point is that the symbols in a computer lack the intentionality (or directedness) that is characteristic of human understanding. Vincent Müller reconstructs the argument with two premises: (1) If a system only performs syntactical manipulation, it cannot acquire meaning. (2) A computer only performs syntactical manipulation. Therefore, Müller has Searle conclude that a computer cannot acquire meaning (Müller, 2024).

Searle's argument is incontestable. If the definition of a computer is limited to syntax manipulation. Then he is correct in stating that it cannot account for meaning.

However, this narrow view of computation as simply syntax manipulation is an incomplete definition of computation. It is a primary misconception underlying Searle's Chinese Room argument. Defining computation by singling out syntactic manipulation is an incomplete account of computation. A complete account of computation must include an account of the meaning associated with the computational system. This is true whether this meaning is intrinsic as in the mind, or offshored in extrinsic agents as is the case with artefactual computers. The misunderstanding that computation can be qualified without meaning (semantics) can be traced back to Claude Shannon's hack of the Analog Principle in computation (Shannon, 1938). Before Shannon, it was fully understood that computation included the physical grounding of meaning into computational items. By outsourcing the grounding of meaning to an external agent, Shannon created the illusion that computation could be fully explained as a process of syntactic movement alone.[44]

This is inaccurate. Syntactic movement without attached meaning is a meaningless causal chain of events. It cannot be defined as computation. Searle himself hints at this with implied absurdity when he says almost any physical system can be described as executing computations. A wall could be said to be performing computations. If we could discern the patterns of molecular movements in a wall, it would surely cohere to some formal structure that can be defined as computation (Searle, 1990). What Searle aims to communicate through absurdity is that a wall is decidedly not a computer. Yet if meaning was applied to its molecular movement in such a way that one could trace input and meaningfully transformed output in its formal structure, it most certainly would be a computer. If one adopts the definition of computation as simply formal structure, then Searle's absurdist stance would be warranted. Almost any physical system can be said to have a formal structure, yet not everything with a formal structure can be a computer. What qualifies computation is grounded meaning. Whether this meaning is imputed on the system by extrinsic agents as in artefactual computing, or is intrinsic as in the mind, grounded meaning is essential to qualify computation.

In early computers, grounding was achieved through the Analog Principle, while in modern computers, meaning resides in the minds of designers who physically embed it through

[44] It's important to note that analog computers also relied on external agents for grounding. The analog models, which give meaning to the computations, had to be designed and built by engineers who determined the meaning the models would impart. However, this approach made it more apparent that meaning grounding is an essential feature of computation. Digital computing obscures this essential role.

conditional statements, and in the mind of users who impute its meaning from shared conventions onto the system. Meaning is an integral part of computing, and any explication of computing is incomplete without accounting for the intentionality represented in computing, wherever this intentionality may reside. This applies even to digital computers which have externalized their intentional content.

One can develop an ‘Insufficiency Reply’, to the Chinese Room argument. The reply suggests that the Chinese Room is an inadequate model of computation because it fails to account for the meaning grounding component of computation. It is a strawman. It presents an insufficient account of computation and critiques it for not doing the precise thing it omitted from the account. In building his case on the idea that a computer only performs syntactic manipulation, Searle fell victim to the Shannon obscuration. Shannon created the impression that the notion of computation could be separated from grounded meaning. By outsourcing the grounding to external agents, he made it seem as though computation could be coherently explained as a process of syntax manipulation alone.

If the Chinese Room was to be upgraded to more sufficiently model the whole computation process; for example, with an analog mechanism that accounts for grounding meaning into syntax to qualify them as symbols, the system will demonstrate some level of understanding. If the person (homunculus) in the room had access to a model that analogously represented the world and linked each symbol they encountered to objects in the model, the symbols would have directedness related to things in the world. The homunculus would realize that these symbols correspond to modelled external objects. Given enough time and memory, they might even learn to respond to the symbols without needing instructions, because they have internalized what the symbols mean and how to react to them. In this scenario, the symbols have underived meaning to the homunculus that relates to features of the modelled external world. The symbols have semantic content.

Searle’s Chinese Room thought experiment is flawed in that it presents an incomplete model of computation. It falls victim to the ‘Claude Shannon obscuration’ which obscures the need for grounding in computation by assigning that responsibility to an external agent. If we adjust the Chinese Room to include the Analog Principle (a key feature of computation before Shannon's work), Searle’s homunculus would begin to understand the symbols it processes. Searle’s errs in modelling computation as purely syntactic processes. Since meaning is an essential part in qualifying computing, it is incorrect to claim that a computer is sufficiently defined by syntactic processing. Formal movements without meaning are simply meaningless causally linked events.

4.5. The AGI agenda and the Analog Principle.

If the mind is an analog mechanism, the AGI agenda can draw lessons on how to model artefactual systems that self-ground semantic properties onto syntactic operations.

To realize any analog mechanism, one needs a physical medium with both continuous and discrete properties that are synchronised. The correlation between the two properties allows information from one property to be translated or transferred onto a second property. This chapter has leaned very heavily on the example of a clock that converts the angular motion of its hands (continuous property) into numerical symbols (discrete property) that can be used to compute the time. Other examples in the same vein are a mercury thermometer and a compass. In all these examples, worldly phenomena are modelled on the continuous property in the form of an analog model that serves to convert structural information onto the discrete properties. Once imbued with meaning from the world the discrete properties become symbols that can be used for computation or combination into more complex forms.

The AP Thesis states that the mind utilizes an analog mechanism in cognition to ground mental computational items. Qualitative states in cognition are an analog model that models signals from the world and encodes structural information from the signals onto computational items hosted on or in the neurophysiological structure of the mind.

The AP Thesis resolves a critical gap in CTM. The CTM has been challenged as an apt analogy of the mind because it could not account for how the mental syntactic processes gain semantic properties, and where these semantic properties fit in the physical causal chain of mental processes. The AP Thesis resolves this issue. It describes the natural mechanism by which underived, non-natural meaning is introduced into mental syntactic operations to ensure that mental syntactic structure is honest to the worldly phenomena it represents.

Early computer architects identified that an analog mechanism was critical for computational grounding. What the early analog computers, like the Harmonic Tide Analyzer were missing in matching the semantic grounding in cognitive computation, was a means by which to sustain an analogous relationship with worldly signals without the aid of extrinsic agency. However, innovation in digital computing advanced in the opposite direction. It served to increase the role of extrinsic agency in computational grounding rather than to remove it. The breakthroughs to digital computing have served to obscure the importance of the Analog Principle in computing and thereby have muddied the analogy between artefactual computing and cognitive computing. The dominant role of extrinsic agency in artefactual computing has served to drive computational approaches further away from the computational self-sufficiency that is mandated for general intelligence.

The AGI engineer ought to look at how cognition leverages qualitative properties in an analog mechanism to self-ground computational items with meaning. The goal of an AGI engineer should be to build a system that can independently receive inputs from the world, model the inputs as analog models with qualitative properties for the purpose of encoding structural information from the model onto computational items.

Computation is the syntactic manipulation of meaningful symbols. All computations require grounded information. By having the ability to self-ground computational symbols, cognitive agents take the first step in ensuring computational self-sufficiency. Artefactual computation relies on engineers to meet this demand. They are not computationally self-sufficient because they need outside help to effect computational processes. For the AGI agenda to be successful, engineers must replicate the functional role of qualitative properties in grounding syntax with meaning.

The next chapter will explore the role of qualitative properties in meeting the positive dictate of computational self-sufficiency. The positive dictate is a demand for intrinsic states and processes that can be credited with self-control.

Chapter 5: On the Umwelt and the Internal Model Principle for self-control.

> *"The nervous system is viewed as a calculating machine capable of modelling or paralleling external events, …this process of paralleling is the basic feature of thought and of explanation."*
>
> The Nature of Explanation (Craik, 1943, p. 121)

5.0. Introduction: The Umwelt and self-control.

When Jakob von Uexküll introduced the concept of the Umwelt in 1909, he most certainly wasn't thinking about the challenge of realizing computational self-sufficiency in artefacts with humanlike general intelligence. The Umwelt refers to an organism's subjective sensory model of the world. It informs an organism's self-control when interacting with a dynamic and complex environment. There are already clear parallels between this idea of a sensory model and the qualitative models discussed in the previous chapter. The previous chapter discussed how an internal model of the qualitative properties of worldly signals supports self-sufficiency in computational grounding. This is crucial in addressing the negative dictate of computational self-sufficiency, which implies a prohibition on creditable external agents in computational processes. In this chapter, I will argue that an internal sensory model much like an Umwelt also addresses the positive dictate of computational self-sufficiency, which demands internal processes and states credited with self-control.

The term 'Umwelt' is a German term that directly translates to environment but more accurately means self-centred world (Kull, 2010). The word has been adopted into ethology and its neighbouring disciplines to mean the unique and subjective sensory world of an organism. The term describes the internal appearance of the world to an organism. The internal appearance of the world is to be understood differently from the environment as the external world, which is the same for all organisms that share the environment (Kull, 2010). The Umwelt is shaped by an organism's subjective sensory perceptions, the external environment, and its biological capabilities and needs. The Umwelt is a subjective model of the world (Kull, 2010). It is the world as it is represented by the faculties of an organism interacting with its native environment.

The Umwelt allows the organism to engage with its environment in a way that reflects its own perceptions, capabilities, and needs. In this sense, the organism acts in its own self-interest, making its decisions based on internal processing of information that relates to its perceptions, capabilities, needs, and what the environment offers to meet those needs. The ratiocination towards actions is driven by the organism's internally accessible information, without any immediate external agency affecting its decision-making process.

The functioning of the Umwelt holds critical lessons for AGI research that is aimed at achieving the Grand Ambition of creating artificial general intelligence that functions comparably to human intelligence. A key feature of this grand ambition is computational self-sufficiency, which includes both a positive and a negative dictate. The negative dictate prohibits reliance on external agents in the computational process, while the positive dictate mandates internal states and processes that can be credited with self-control.

Mainstream current approaches to artefactual computation fail to meet both dictates of computational self-sufficiency. These systems rely on external agents not only to ground symbols but also to define the problems they are meant to solve. Because external agents play such a central role in grounding meaning and problem-solving, they take on most of the responsibility and credit for the system's problem solving.

A more effective approach to achieving the Grand Ambition would be the replication of the functional roles in the human mind. By replicating the principles that enable the mind to exert self-control, we could bring computational models closer to the way the mind functions. The goal is to identify the processes through which the mind exerts self-control and determine whether these can be replicated in AGI models to enable self-directed problem-solving in dynamic environments.

This is where the concept of the Umwelt becomes important to AGI research. The Umwelt suggests that the mind leverages an internal sensory model that integrates information about the world, the organism's internal states, and its needs to the end of directing self-control.

The idea of the Umwelt as an internal model that informs self-control parallels the concept of a controller in the Internal Model Principle (IMP) of Control Theory. In the IMP, a controller is an internal model that regulates the relationship between a system's inputs and outputs to optimize performance in changing conditions. This is especially useful in systems that face broad and dynamic inputs. The controller governs processing to produce the appropriate output in response to environmental and internal dynamics. Simply put, a well-designed controller exerts control over system processing, enabling it to behave appropriately in response to environmental and internal changes.

This chapter will argue that the qualitative experiencing that gives rise to an Umwelt implies a sense of an emergent internal entity, often referred to as the "inner self," which is credited with self-control. It will further contend that the inner self functions similarly to how an internal model in control theory acts as a controller that governs the system's self-control.

It's important to clarify that the articulation of the 'inner self' that is discussed in this chapter differs in context from the typical discourse on the self in the philosophy of AI. The primary discourse on the self in the philosophy of AI tends to focus on whether AIs, which do not appear to have a cohesive sense of self as a continuous personal identity, can have moral agency and be treated as moral patients. While this research falls within the philosophy of AI, this articulation of the self is not the focus of this discussion. The use of the term "inner" as a prefix in the word 'inner self' is meant to distinguish the discussion of self in this thesis from the main discussions of the self in the wider philosophy of AI debate. This chapter will focus on the inner self as the author of self-controlled action, rather than the self as the locus of moral responsibility.

The chapter is structured in two parts. The first section (5.1) will explore how the sense of an emergent inner self that is credited with exerting self-control, is implied by qualitative experiencing. From these implications, the section will abduce the functional roles of the inner self to cognition. The second section (5.2) will introduce the Internal Model Principle (IMP) from control theory and argue that the internal model in this principle serves functional roles that are analogous to those of an inner self in cognition. The section will conclude that the IMP can be adopted as a physicalist accounting of the functioning of the inner self in effecting self-control in cognition. The final section (5.3) will conclude by highlighting how the Internal Model Principle can support the type of computational self-sufficiency that is interesting to the AGI agenda.

5.1. The inner self as emergent from qualitative experiencing.

5.1.1. The inner self as the author of action.

The self-control that defines the mind is often credited to an emergent inner self that is viewed as the author of action and behaviour. This is evident in the language that is used to reference the type of self-control that the mind affords. The prefix 'self' in the words self-control, self-determination, self-propelled all signal that the agent defined as controlled, determined or propelled is such that these actions are wholly creditable to an emergent inner self within the agent. This outlook goes beyond the way we use language. It appears to conform to an intuitive understanding of the mind. It seems apt to think that the mind gives rise to an inner self that is the subject of experiences and is credited with self-control.

The danger of this outlook is that it might lead one to what Daniel Dennet calls Cartesian Materialism (Dennett, 1991). It creates the impression of the inner self as a homunculus and the mind as a Cartesian theatre (p.107). In this outlook the inner self views all incoming sensory experiences and processes control information for the agent.

The obvious flaw in this outlook is that it implies an infinite regress of inner selves. If the inner self functions by experiencing information and processing output for the agent, then there must be an inner-inner self that experiences the experienced information and processes the inner self's output, and so on *ad infinitum*. This is obviously not a satisfying explanation of cognitive self-control.

The idea of an inner self in the form of a *Dennetian* homunculus or a Cartesian theatre is clearly wrong. But none the less the inner self remains a forceful intuition. Something about how the mind functions makes it appear that there is an inner self that is closely related to experiencing the world and is credited with control and decision making. What is needed for a process-centric approach to AGI that aims to replicate the processes of the mind is a naturalist account of how this inner self-controller emerges from the processes of the mind and how it affects or effects decision making.

5.1.2. The concepts of Self NOT covered in this discussion.

The interest of this thesis in the concept of self is very narrow. The interest is limited to an emergent sense of an inner self that is credited with experiencing and authoring actions. The goal of the discussion is to understand the processes that might give rise to a sense of centred and subjective self-control. This interest in the inner self as the subject of experiencing and the author of action in this thesis should be separated from the broader debates about the self in philosophy. It is worth saying a few words about the breadth of conceptualizing around the self that will NOT be discussed in this thesis.

The self is a nebulous concept and a broad topic of discussion in philosophy. The multifaceted and complex nature of the self was immediately apparent to theorists engaged in its study. In one of the earliest philosophical works dedicated to theorizing about the self, William James categorized the various facets of self into the physical self, mental self, spiritual self, and the ego (James, 1890). Since James' offering, the categories have been considerably supplemented. In Galen Strawson's the Self and the SESMET (Strawson, 1999), Strawson surveyed an inexhausted list of categories of self that included,

> *"…the cognitive self, the conceptual self, the contextualized self, the core self, the dialogic self, the ecological self, the embodied self, the emergent self, the empirical self, the existential self, the extended self, the fictional self, the full-grown self, the interpersonal self, the material self, the narrative self, the philosophical self, the physical self, the private self, the representational self, the rock bottom essential self, the semiotic self, the social self, the transparent self, and the verbal self* (Strawson, 1999, p. 100)*."*

The different concepts of 'the self' have been the subject of theorizing in different philosophical traditions. An incomplete list of how various traditions approach the self includes essentialism, which sees the self as an unchanging essence or soul. Empiricism which views the self as a collection of experiences. Existentialism describes the self as a dynamic shaped by actions. Phenomenology focuses on the self as first-person lived experience. Psychoanalysis sees the self as fragmented in being split into conscious and unconscious parts. Social constructivism understands the self as shaped by society and culture. While postmodernism sees it as decentred and fluid.

While I recognize that all these other discussions of the self are important, the interest in this discussion is limited to the narrow conceptualization of the inner self that is seen as the subject of experience and the author of actions (Strawson, 2011, p. 253), (Frege, 1956, p. 299). With this narrow focus, it is apparent that the different centres of gravity that ground all the above concepts and discussions of a self could not all possibly be of relevance to the agenda pursued in this thesis.

However, even with this narrowed focus on the aspects of an inner self as the subject and author of self-control, the inner self remains an incredibly varied and complex concept. The inner self is what experiences qualitative states and it is what is credited with responding to their prompts. As discussed in prior chapters the experience of qualitative states can be varied and incredibly complex which makes the inner self a difficult subject to unpack. For example, humans, experience a range of qualitative states. These include perceptual states like seeing and hearing, bodily sensations like pain and hunger, reflective states like thinking and planning, emotional states like anger and jealousy, etc. All these qualitative states appear to be of different types and appear to engender actions in different ways. Attempting to understand the concept of the inner self that emerges from these varied qualitative states and their roles in engendering action is potentially confusing.

The ideal approach is to reduce these qualitative states by stripping away all their complexity and reducing them to their most essential feature. What is sought is the self that is the subject of the most basic account of qualitative experiencing. Fortunately, the work of delineating a minimal sense of qualitative experiencing has been done. The concept and the agenda have been appropriately named the 'minimal self.'

5.1.3. The Minimal Self as the most basic conceptualization of the Inner Self.

The minimal self describes the concept of an inner self that is the subject of the most basic and immediate qualitative experience. The minimal self has been stripped of all its unessential features to leave only that which experiences what is termed as immediate self-consciousness (Gallagher, 2000), (Gallagher & Zahavi, 2012). Immediate self-consciousness occurs prior to any reflection on consciousness and therefore before any other mental activity. It is a pre-reflective, first order phenomenal experience, and it is the essential foundation on which all other mental experiences are built on.

There is broad consensus in phenomenology that all experiential states stripped to their essential nature are reducible to a pre-reflective, immediate self-consciousness. Edmund Husserl called this immediate experience, a self-appearance, (Husserl, 2019) Michel Henry called it a self-manifesting experience (Henry, 1973), Maurice Merleau-Ponty called it a self-givenness (Merleau-Ponty, 2012), Jean-Paul Sartre argued that immediate self-consciousness is not simply a quality of experience, but the very mode of experience (Sartre, 1956). In all these descriptions there is an implied something at the receiving end of immediate self-consciousness. Something is experiencing Husserl's self-appearance, Henry's self-manifestation, Merleau-Ponty's self-givenness, and Sartre's mode of experience. This something is a minimal self. An inner self as it would appear if it has been stripped of all its higher order roles in cognition.

The minimal self simplifies the concept of the inner self by stripping away all its higher order and complex elements and reduces the concept to the easily comprehensible aspect of experiencing. A theory of the minimal self holds that the varied and complex types of mental experiences like perceptual experiences, reflective experiences, emotional experiences, etc, all have at their root in an immediate, pre-reflective qualitative quality, without which there will be nothing it is like to experience the mental states (Zahavi, 2002). The concept of the minimal self therefore affords a simple and neat concept to work with. Rather than get bogged down with the complex details of trying to understand the concept of the inner self from the diverse types of experiential mental states, the concept of minimal self allows a focus on the very aspect of hosting an experience in whichever form.

The focus on the minimal self offers another distinct advantage. It is a simple enough concept that it can apply to all types of minds. If we move beyond an anthropocentric perspective, we recognize that mental states occur in many creatures, not just humans. While all creatures with minds have some form of experiential states, it is reasonable to assume that not all of these states are the same across species. By centring the minimal self, the emphasis shifts to the act of experiencing itself, rather than the specific nature of those experiences. This approach sets aside debates about the types of experiences and instead focuses on the fundamental, immediate self-experiences that can be attributed to all minds..

5.1.4. The Affordances of an inner self.

So, what exactly is an inner self? The simplest approach to this question is to ask it of the inner self's minimal form that is the subject of immediate experiential mental states. One way to answer the question is to attempt a definition that locates some form of ontic structure on which to ground the concept. Galen Strawson offered up a broad framework intended to establish enough common ground to accommodate broad perspectives on a definition of self. He stated that the inner self is conceived or experienced as "(1) a thing, in some sense, (2) a mental thing, a single thing that is single both (3) synchronically considered and (4) diachronically considered, (5) a thing that is ontically distinct from all other things, (6) a subject of experience and (7) an agent and (8) something that has a certain personality (Strawson, 1997, p. 4)." Strawson is aware that his proposal might appear vague, contain redundancies, and feature entailment relations between some of the elements. His interest is in providing a framework that accommodates diverse perspectives on the self.

Yet, even with the concession that Strawson's description is intended to be broad, his characterization of the inner self will leave many unsatisfied. For example, Daniel Dennet believes the inner self is an abstraction of a collection of fragmented phenomenal experiences and states (Dennett, 1992). Wolfgang Prinz argues that the inner self is a construct imported from others by way of phenomenal representation of others and the world (Prinz, 2019). While Aaron Sloman argues that the inner self is inseparable from the agent itself, any attempts to define an inner self as an ontically distinct thing arise from a misunderstanding of reflexive language (Sloman, 2021). A definition of inner self as a real entity is bound to draw broad and vigorous disagreements.

One path through this metaphysical impasse is to ground an understanding of the inner self on what it affords. Definitions grounded in an affordance define a thing by a perceived quality or property that makes clear how a thing can be or should be used. Affordances are regularly used to ground definitions. For example, a definition of the word 'bed' is a piece of furniture on which to sleep or rest.[52] Grounding definitions on affordances is not limited to artefacts. The approach is widely used and accepted in the natural sciences. For example, an ear is an organ for hearing and equilibrium. Gravity is a force by which planets are attracted to each other and kept in orbit. In all these examples, the affordance of the entity is the salient part of its description.

Does the inner self offer any concept defining affordances? This is an especially intriguing approach for an AGI researcher. This is because AGI research is interested in mental states and processes from the perspective of functional roles and not their form. The inner self is interesting to an AGI researcher in so far that it affords something to cognition.

The inner self affords two things to cognition. First it creates a sense of body ownership. This can be defined as a feeling of ownership over the agent that experiences, and a recognition of

[52] (Merriam-Webster Dictionary, 2022)

the entity's embeddedness in a broader environment (Gallagher, 2000), (Braun, et al., 2018), (Hafner, et al., 2020). Secondly it creates a sense of agency. This is a feeling of being able to initiate and control some key cognitive actions, and through this exert some control on the agent's interaction with the environment in which it is embedded (Gallagher, 2000), (Haggard, 2017), (Braun, et al., 2018), (Hafner, et al., 2020).

The two affordances of the inner self are so closely intertwined that at first pass might appear as the same thing. When I reach out and knock over a vase, I have an instinctual awareness that this is my action. This creates the impression that the ownership of a body's action is the same as agency. For example, one might be led to believe that the owner of a body's action is the same person who caused the action. However, the two are separable. For example, imagine that someone pushed my hand and caused me to knock over the vase. In this case I might have a sense of ownership over the moving hand, in that, I recognize that it is my hand that moved to knock over the vase. However, I can deny agency over the movement, in the sense that I deny responsibility for the cause of the movement. Involuntary actions such as described in this example are a good way to conceptualize how the sense of body ownership and the sense of agency are separable.

The two affordances, the sense of body ownership and the sense of agency, are affordances offered by an inner self that define its role in cognition as the creditable entity in self-control.

A definition of inner self that centres its affordances is a less divisive issue than the question of the inner self's actual form. In fact, one can trace a consensus on the affordances of an inner self even in wildly differing descriptions of the nature of an inner self. Take for example the brief survey of theories on the form of inner self above. While there is real disagreement on the specific form of an inner self, one can extract common ground on the affordances offered by an inner self. Strawson's 6th and 7th point in his broad framework of self suggest that an inner self is creditable as a subject that owns experiences and has agency (Strawson, 1999). Prinz thinks that the inner self is a construct imported through the action perception of others. Prinz therefore wouldn't dispute that the inner self affords a sense of body ownership and agency, he just thinks that we understand these affordances as constitutive of a concept of an inner self from watching others exercise ownership and agency over their body and actions (Prinz, 2019). Even Dennett and Sloman who think that the inner self is not a real thing will find something to agree about in the affordances of an inner self. Dennet recognizes that even as a fictional item, the inner self is useful for creating a bound entity that experiences and responds to experience (Dennett, 1992). While Sloman's real problem with conceptualizations of an inner self is that it is conveyed as something that is ontically distinct from personhood. Sloman does not dispute that there is something that has a sense of body ownership and agency. He however argues that this it is the person as a whole. An inner self is simply an illusion to which these affordances are attributed to because of the improper use of language (Sloman, 2021).

In whichever way the inner self is conceptualized, there appears to be, at best a consensus, or at worst, no grounds for objecting that the inner self will afford a sense of body ownership and agency.

5.1.5. The role of qualitative experiencing in realizing the affordances of an inner self.

A careful rumination on our intuitions of an inner self will reveal it is intimately related to qualitative experiencing. The inner self is the subject that experiences qualitative states and is credited with the actions they engender. The definitive parts of the inner self, the sense of body ownership and a sense of agency, appear to emerge from qualitative experiencing. The sense of body ownership is a sense of *mineness* for one's body parts, feelings, and thoughts (Gallagher, 2000), (de Vignemont, 2011), (Tsakiris, 2016). The sense of agency is the sense of authorship over actions (David, et al., 2008), (Moore, 2016), (Bayne & Pacherie, 2007).

How do the sense of body ownership and agency arise from qualitative experiencing?

To answer this question, one must return the focus to the minimal self. As a reminder the minimal self is the conceptualized subject of immediate experiencing. An interest in the minimal self is an interest in the most fundamental aspects of experiencing. If the sense of body ownership and the sense of agency are emergent within a minimal self, it means that they are irreducible features of experiential states. To put the same point in other words; If the sense of body ownership and the sense of agency can be located in the minimal self, then it would mean that the very act of having an experiential state bestows these affordances on the bearer of experience. This position demands an explanation.

Most accounts of the minimal self has conceptualized the minimal self as being constituted of both the sense of agency and the sense of body ownership based primarily on appeal to psychological experiments that claim to demonstrate empirical evidence for both senses (Gallagher 2000), (de Vignemont, 2011), (Tsakiris, 2016), (Forch & Hamker, 2021), (Blanke & Metzinger, 2009), (Newen & Vogeley, 2004), (Hafner, et al., 2020), (Braun, et al., 2018). However, there is room for an inductive account for why having any form of experiential state implies having a sense of body ownership and a sense of agency.

The next section will trace how the affordances of a sense of body ownership and a sense of agency are implied by immediate qualitative experiencing and abduce from this account the functional roles of these emergent senses.

5.1.6. The Implications of immediate qualitative experiencing.

Before explaining how the sense of body ownership and the sense of agency are implied by immediate qualitative experiences, it's important to offer a disclaimer. The subjective nature of qualitative experiences makes it difficult to draw universal conclusions about their nature and implications. Since we can only access these states through personal, first-hand experience, any attempt to describe or analyze them as a universal law is inherently inductive. It involves making inferences from individual experiences to broader principles. The broader these inferences, the less justifiable they become. For instance, it may be reasonable to argue that insights from one person's qualitative experience could apply to other humans. However, this justification weakens as we move further away from human experiences to those of other kinds of minds.

That said, inferences drawn from human subjective experiences don’t need to apply universally to all minds to be useful for AGI research. The focus of AGI is on understanding the functioning of a mind and whether its functional aspects can be replicated in artificial systems. For this purpose, it is sufficient to draw conclusions from one type of mental functioning, even if there are many types. To illustrate, imagine an engineer building a house. While there are many styles and approaches, the engineer only needs to replicate one successful design to achieve their goal. They don't need a universal law governing all house-building methods. Similarly, an AGI engineer aiming to replicate the mind’s functioning only needs principles derived from one successful example, not from a universal law that applies to all minds. These principles need only demonstrate functionality in one form of cognition to be useful.

With this disclaimer registred it should be noted that the implications of immediate qualitative experiencing that give rise to the sense of body ownership and the sense of agency that are discussed here, are inferred from human experience. It is enough for our agenda that these implications are true to at least one form of cognition.

Qualitative experiencing and the Sense of body ownership.

The sense of body ownership is a pre-reflective experience or sense that the experiencer owns its body and its movements (Gallagher, 2000). This pre-reflective experience can be inductively derived from at least three consequent implications of having an immediate qualitative experience.

An immediate qualitative experience will imply -

1. **Demarcation of the Body.**

The primary implication of having an immediate qualitative experience is the demarcation of the body as the host of that experience. In other words, having a qualitative experience implies that there is a bearer of that experience (Frege, 1956, p. 299), (Strawson, 2011). A qualitative experience is manifested as a bodily sensation that has distinct qualitative properties. The sensation is discernible, differentiable, continuous, and potentially variable in intensity. Bodily sensations highlight specific parts of the body that host the sensation, marking them as belonging to the experiencer. An immediate qualitative experience makes the experiencer experience the sensation as felt within a body that is both accessible to them and separable from the external world.

It's important to clarify that this sense of body demarcation should not be confused with a requirement for a strictly localized embodiment. A sense of body ownership can be differentiated from embodiment (de Vignemont, 2011, p. 84). While a qualitative experience demarcates the bearer of the experience as separate from the rest of the world, it does not demand that the body must be a single, continuous mass. Psychological experiments like the Rubber Hand Illusion demonstrate that a sense of bodily sensations can extend to include non-extended parts (Botvinick & Cohen, 1998). In this experiment, participants perceive a rubber hand, a non-extended object, as part of their own body.

However, these experiments also support the idea that qualitative experiencing is tied to a sense of body ownership. They reinforce the link between experiencing bodily sensations and the demarcation of the perceived seat of those sensations as belonging to the experiencer. In the case of the rubber hand illusion, the sensation leads the subject to treat the rubber hand as an accessible part of their own body, showing that the experience can cause the experiencer to incorporate the ascribed seat of sensation into their sense of body.

2. Situatedness.

The second key implication of having an immediate qualitative experience is the sense of being situated in a specific environment. When the body is demarcated as distinct, it becomes accessible and recognizable as separate from its surroundings. The environment is perceived as foreign to this demarcated body. However, while the environment is external, it is still recognized as the source of signals that provide qualitative sensory information.

The demarcated body receives information from the environment in the form of worldly stimuli. This information is represented as qualitative input, with properties that meaningfully reflect the characteristics of the external stimuli. In other words, these qualitative states mirror the properties of the world's stimuli in matching their discernibility, differentiability, continuity, and variableness in some meaningful way. This creates the sense that the demarcated body is embedded in an environment that continually provides it with qualitative information.

3. Perspectivalness.

Thirdly, an immediate qualitative experience implies a centre of perspective. The previous points established that qualitative experiencing implies a sense of the demarcation of the body, distinguishing it from the surrounding environment, while also recognizing that the environment feeds the body sensory information. The awareness of these sensations from the environment is called perception. The centre of perspective refers to the locus of perceptual experience. It is from this centre that the subject perceives stimuli or dynamics in the world that cause the sensory information. In other words, perceptual qualitative states manifest as first-person perspectives on the sources of sensory input.

The three implications of an immediate qualitative experience; body demarcation, situatedness, and perspectivalness, together form the emergent sense of body ownership. This sense of body ownership presents as a demarcation of the cognitive entity from the environment in which it is embedded and a first-person perspective for the cognitive entity onto the world.

Qualitative experiencing and the Sense of Agency.

There have been two broad approaches for accounting for a sense of agency (Bayne & Pacherie, 2007), (Gallagher, 2007). These include a top-down account in which a sense of agency is derived from a reflective narrative account of self (Dennett, 1992), (Gallagher, 2020). And a bottom-up approach that locates a sense of agency in neural processes that relate or respond to sensory signals (Gallagher, 2000), (Blakemore & Frith, 2003).

Our interest being in the minimal self that is the subject of immediate experiencing, precludes an interest in the top-down account of a sense of agency. This is because a reflective narrative account implies a higher order reflection in constructing a narrative self. Our interest is more atomistic. We want to understand how a minimal self and its constituent sense of agency is implied by immediate pre-reflexive experiencing.

The bottom-up approach to the sense of agency locates the sense of agency as emergent from neural responses to qualitative sensory signals. There are two immediate implications that must precede any form of response to a qualitative sensory experience.

1. Discernment and differentiability.

Having an immediate qualitative experience implies the ability to both discern and differentiate the qualitative experience. A qualitative experience is defined by its discernibility and its form. This implies that discernibility and differentiability are fundamental properties of a qualitative experience. To register an immediate experience, the experience must be recognizable as being distinct from having no experience. Additionally, it must be differentiable from other types of experiences.

2. **Conditions to exercise choice.**

The ability to discern and differentiate between qualitative experience implies the conditions to exercise choice. This point can be more clearly understood from its converse position. In the absence of discernible differences, there is no basis on which to exercise choice. For instance, in conditions of perfect uniformity, where there are no distinguishable features, there are no grounds for selecting one option over another. The ability to make choices depends on the presence of discernible and differentiable properties between the available options. When an experience can be distinguished and differentiated, it establishes the conditions for exercising choice. Which is fundamental for exercising choice in generating a response to a signal.

Operationalizing choice is supported by a different type of qualitative experience. Hedonic signals, such as feelings of pleasure or displeasure, guide the experiencer in making choices in response to their environment. These choices are based on preferences for the quality of one experience over another. In other words, the experiencer might prefer one type of experience over another for various reasons. However, for hedonic preferences to drive choice, the entity must first be able to discern and differentiate between qualitative experiences.

The key takeaway is that the ability to discern and differentiate distinct experiences creates the conditions necessary for making choices between them. Hosting qualitative experiences, which are both discernible and differentiable, enables the exercise of choice. These choices are then later motivated or operationalized by qualitative preferences or other influencing rewards.

Immediate qualitative experience implies an atomistic sense of agency. The ability to discern and differentiate between qualitative experiences, along with the resulting capacity to choose between them, gives rise to a sense that the experiencer has the basis for exercising agency.

5.1.7. The functional roles of the inner self.

By understanding how qualitative experiencing implies a minimal self, we can abduce the functional roles of the minimal self to cognition. This process involves working backwards from the implications of qualitative experiencing to identify how the implied senses of body ownership and agency contribute to the cognitive process.

By working backwards from the three implications of qualitative experiencing that lead to a sense of body ownership, we can infer the following functional roles of the sense of body ownership:

- **Demarcation of body**: This role creates a delineation for the body and tracks its internal dynamics.
- **Situatedness**: This role makes the experiencer aware of the environment in which the body is located and tracks its changing dynamics.
- **Perspectivalness**: This role centralizes all the gathered information for processing and centres an awareness of the body in relation to the environment.

Together, these functional roles describe a **body schema**, which refers to the representation of one's body, its position within the environment, and its orientation relative to that environment (de Vignemont, et al., 2021), (Graziano & Botvinick, 2002).

Table 5.1. The functional roles of a sense of body ownership.

Implication of qualitative experiencing	Functional role in cognition
Demarcation of Body	Create a delineation of the body and track its interoceptive information.
Situatedness	Create awareness of environment and track shifting dynamics from environment.
Perspectivalness	Create a locus for interoceptive and exteroceptive information and centres awareness of body position in environment.

The two implications of qualitative experiencing that give rise to the atomic sense of agency form the foundations for self-control:

- **Discernability and differentiability:** This role involves differentiating incoming information based on its qualitative features.
- **Conditions for exercising choice:** These arise from the ability to discern and differentiate between phenomenal experiences.

The ability to discern and differentiate qualitative experiences, set the conditions to exercise motivated choice. The conditions to exercise choice are a key factor that underpins self-control.

Table 5.2. The functional roles that yield a sense of agency.

Implication of qualitative experiencing	Functional role in cognition
Discernability & differentiability	Ability to discern and differentiate between different signals.
Conditions for exercising choice.	Create the conditions to exercise decisions based on discerned differences.

Earlier, it was suggested that something about how the mind functions gives the impression of an inner self, that experiences the world and is credited with control and decision-making. By showing how the implications of qualitative experiencing lead to a sense of body ownership and agency, we can begin to understand the origins of this intuition.

However, we can go a step further. Through extrapolation, the implications that give rise to the sense of body ownership and the sense of agency have been interpreted as functional roles that define a body schema and establish grounds for self-control. These functional roles can be used to support a physical account of how an emergent inner self exerts self-control.

The next section will explore how the emergent inner self has functions that are like a controller in control theory. Control theory defines physical principles for self-control.

5.2. The Internal Model Principle.

5.2.1. The Controller in Control Theory.

Control theory in engineering focuses on managing dynamic systems in machines and engineered processes. Its goal is to develop controllers in the form of internal states, models, or algorithms that govern input processing to ensure the system's stability and optimal output. Control theory provides valuable insights into self-control within systems influenced by changing forces and inputs.

A key concept from control theory that is relevant to understanding how the inner self functions as a controller is the Internal Model Principle. This principle explains how to design self-control mechanisms in systems subject to varying inputs and forces. It states that to achieve effective self-control, a system must incorporate an internal model that represents the dynamics influencing control decisions (Wonham, 2018). This internal model monitors both internal and external changes to inform control decisions (Camacho & Bourdons, 2004).

The principle originated with the Good Regulator Theorem, which proposed that an effective regulator must model the system it controls (Conant & Ashby, 1970). Later developments refined this idea, suggesting that a control system should simulate the system it manages, enabling it to predict disturbances to make appropriate adjustments (Francis & Wonham, 1976). In simple terms, an internal model collects information about external factors and internal system states, to generate and issue control instructions. In more advanced systems, the internal model also monitors feedback to correct errors continuously.

The Internal Model Principle is particularly interesting to understanding the emergent inner self, which is credited with exerting self-control in response to a dynamic environment. A system designed using this principle can adapt to changes in both its environment and in it-self. This mirrors how we understand the role of the inner self as creditable with self-control, and it mirrors the functions inferred from the inner self in representing a body schema for the ends of exercising self-control.

5.2.2. The Internal Model as analogous to the inner self.

The controller in the Internal Model Principle has functional roles that closely mirror those we identified as arising from qualitative experiencing in cognition. In the previous section, we divided the functional roles that are implied by qualitative experiencing into two groups; those that contribute to the emergent sense of body ownership and those that contribute to the atomistic sense of agency. These two senses, body ownership and agency, not only characterize the minimal self but also define it in terms of its cognitive functions.

The controller in IMP as functionally analogous to a sense of body ownership.

The sense of body ownership is an emergent feeling of ownership over one's body, its experiences, and its actions, and a recognition of its embeddedness in an environment (Gallagher, 2000), (Braun, et al., 2018), (Hafner, et al., 2020). This sense emerges from three key implications of a qualitative experiential state:

- Demarcation of the body
- Situatedness in an environment
- Perspectivalness

By working backwards from these implications, the functional roles of the sense of body ownership in cognition can be inferred as follows:

- Creating a delineation of the body and tracking interoceptive information (internal body states)
- Creating awareness of the body's situatedness in an environment and tracking dynamics within that environment

- Centralizing interoceptive (internal) and exteroceptive (external) information for processing, and centring awareness of the body's position relative to its environment

These functional roles have been described as giving rise to a body schema. The concept of a body schema has been adopted in engineering for both descriptive and directive roles (Graziano, 2017), (Hoffman, et al., 2010). The internal model described in the Internal Model Principle is like a body schema that delineates the body of the system and directs action. According to the Internal Model Principle, the internal model delineates the system's spatial properties (body), monitors internal and external changes, and centralizes information for decision-making.

In this way, the abduced functional roles of the sense of body ownership align closely with the roles outlined by an internal model in the Internal Model Principle.

The IMP as functionally analogous to an atomistic sense of agency.

The sense of agency is an emergent feeling of being able to exercise choice over some cognitive actions, and through this exert some control on the cognitive entity's interaction with the environment in which it is embedded (Gallagher, 2000), (Haggard, 2017), (Braun, et al., 2018), (Hafner, et al., 2020). Our interest in this section is in the bottom-up atomistic sense of agency rather than the top-down narrative sense of agency. This atomistic sense of agency has its roots in two key implications of qualitative experiencing:

- Discernibility and differentiability
- Conditions for exercising choice

These implications can be inferred as definitive of two functional roles of the atomistic sense of agency:

- The ability to discern and differentiate between interoceptive (internal) and exteroceptive (external) signals
- The creation of conditions to make motivated choices or decisions based on these discerned differences

There are clear parallels between the inferred functional role of the atomistic sense of agency and the principles outlined in the Internal Model Principle. In control theory, the internal model processes various interoceptive and exteroceptive signals, each represented in the internal model as distinct and identifiable features. This ability to differentiate between signals enables the system to make decisions based on the advantages or disadvantages of the detected signals. It is important to note that the primary function of the internal model is to support control processes. Its design ensures that it facilitates control in light or in spite of shifting dynamics. In this way, the internal model in control theory serves a similar role to the emergent sense of agency in cognition, both establish the conditions for self-control in dynamic systems.

The internal model in the Internal Model Principle offers similar functions to the sense of body ownership and the atomistic sense of agency in cognition. The inner self, which emerges as an internal body schema from qualitative experiences, parallels the role of an internal model in facilitating self-control. There is emerging evidence from the field of robotics that the two are similar. Vaughan and Zuluaga offered one of the first proof of concept that an internal model of a system and its environment allow a robot to self-control actions in complex environments (Vaughan & Zuluaga, 2006). Bongard, et al. describe a 4-legged starfish like robot that makes use of an internal model, both to enable the robot to learn its own body morphology and self-control for the purpose of shedding light on self-modelling in animals (Bongard, et al., 2006). Marques and Holland argue that an internal model in the form of a simulation can support the steering of behaviour in robots by giving the robot a functional imagination (Marques & Holland, 2009). The similarities between the abduced roles of an emergent sense of an inner self and the internal model in the IMP in demarcating body and effecting self-control suggest they operate on similar principles, even though there are some differences in representation and complexity (Hoffman, et al., 2010).

Understanding the emergent inner self as functioning like an internal model aligns with our intuition that there is an emergent inner entity in cognition that is closely tied to experiencing the world and is responsible for control and decision-making. The Internal Model Principle provides a more compelling explanation for this intuition than imagining the inner self as a homunculus in a Cartesian theatre as Dennett suggests sardonically. It also offers a physical account for how this intuitive inner self can influence and execute control.

5.2.3. The Internal Model Principle and AGI research.

Although the Internal Model Principle was originally formulated in engineering, its creators view it as a universal physical principle. Roger Conant and Ross Ashby stated that their Good Regulator Theorem applies to all regulating and self-regulating homeostatic systems (Conant & Ashby, 1970). Bruce Francis and Walter Wonham, who refined the Good Regulator Theorem into the Internal Model Principle, also saw it as relevant for motor control in organic systems (Francis & Wonham, 1976). Wonham later emphasized that the Internal Model Principle in engineering mimics internal models in human experience (Wonham, 2018).

Cognitive science similarly acknowledges the concept of the body schema as a sensorimotor representation of the body that informs action (Ataria, et al., 2021). While there are some variations in how the body schema is conceptualized, there is an emerging consensus that it represents a range of sensorimotor body representations based primarily on input information. This understanding is supported by empirical studies (Gallagher, 1986), (de Vignemont, 2010), (Maravita, et al., 2003).

There is also work in philosophy in support of the idea of the inner self as an internal qualitative model for controlling actions. Philosopher Kenneth Craik proposed that the nervous system should be seen as a calculating machine capable of modelling reality internally, with this internal model serving as the foundation of thought and explanation (Craik, 1943). Craik described organisms as possessing an emergent small-scale model of external reality and of their own possible actions within their head, that enabled them to try out various alternatives, to decided which actions are the best of them, and to react in a much fuller, safer, and more competent manner to the emergencies which they face. He compared mental models to scientific models, both of which represent natural phenomena and help predict outcomes (Craik, 1943). Building on this idea, Philip Johnson-Laird developed the mental model theory, suggesting that humans use mental models to understand the world, draw inferences, simulate outcomes, and predict events (Johnson-Laird, 2004).

If the evidence from control theory, cognitive science's body schema theory, and mental model theory provides insights on a physical account of how the mind governs self-control, then these insights hold valuable lessons for the AGI agenda. AGI, in its Grand Ambition, seeks to develop intelligence that is functionally comparable to human intelligence.This requires internal processes or states capable of self-control. A computational system designed with processes based on the Internal Model Principle would bring it closer to a model of general intelligence that is functionally comparable to human intelligence.

It should be noted that internal models, or body schemas are not a novel concept in artefactual computing. They have already been proposed and used in robotics and computational engineering to simulate robot's environments and the robot's themselves, to help the robots to anticipate, predict, and plan their actions (Blum, et al., 2018), (Weinstein, et al., 2024) (Winfield, 2014), (Hoffman, et al., 2010).

There is even evidence to suggest that there is an overt awareness that the proposed internal models in engineering projects serve the role of a self for the robots. John Holland discusses using an internal model to account for a robot's mind (Holland, 2004). Alan Winfield states that an internal model of self accounts to some degree for a self when he writes that

> *"Does having an internal model make a robot self-aware? The answer to this question depends of course on what we mean by 'self-aware'. But, in some straightforward sense, if a robot has an internal model of itself, then that model accounts for the self in self-aware* (Winfield, 2014, p. 238).*"*

The evidence supporting the use of an internal model in robotics for enabling self-control validates the proposal that the Internal Model Principle (IMP) is interesting to AGI research. It's important to note that the AGI concept discussed in this thesis aims to create systems that function comparably to human intelligence. To achieve this, the thesis identified computational self-sufficiency as a key requirement. Human intelligence is computationally self-sufficient. Computational self-sufficiency has two aspects: a positive and a negative dictate. The IMP can be critical addressing the positive dictate, which calls for internal states and processes that can be credited with self-control.

However, fully addressing computational self-sufficiency also requires tackling the negative dictate. This dictate prohibits reliance on external agents in problem-solving. As highlighted in the previous chapter, current computational systems, including the robots mentioned here, still depend on external designers to ground information and calibrate the logic and instructions for computation. This gives these designers outsized credit for problem-solving. This is a departure from how the mind works.

For AGI systems striving toward human-like intelligence, the challenge is to combine both the Analog Principle from the previous chapter and the IMP discussed here. This combination should enable a system to achieve computational self-sufficiency in generating self-control in a way that convincingly demonstrates the independence and fully accountable decision-making abilities of the mind. In the next chapter we shall attempt to demonstrate how the two can be combined in a model of computational self-sufficiency.

5.3. Conclusion: The Umwelt and the Internal Model Principle.

Jakob von Uexküll’s concept of the Umwelt describes a subjective sensory model of the world that helps an organism exercise self-control as it navigates and interacts with a dynamic, complex environment. This idea aligns well with theories in engineering, cognitive science, and philosophy that explain how an internal model can enable self-control over actions.

The concept of the Umwelt is particularly relevant to AGI research, especially in the pursuit of artificial general intelligence that functions comparably to human intelligence. A key feature of this Grand Ambition for AGI is computational self-sufficiency that requires internal states and processes creditable with self-control. The AGI agenda should seek computational self-control that is analogous to the self-control that characterises cognitive computation. The AGI agenda can learn from how an Umwelt supports the emergence of a sense of body ownership and a sense of agency, both of which are leveraged to inform controlled actions.

This chapter has argued that the sense of body ownership and the sense of agency function quite similarly to the roles effected by an internal model in the Internal Model Principle to effect self-control. The next chapter will explore how both the Analog Principle (from the previous chapter) and the Internal Model Principle can be applied to computational models to support intentional actions in response to challenges in a dynamic environment.

It will also examine whether models enhanced with these principles truly achieve the type of computational self-sufficiency that can be credited with the intentional actions that are the hallmark of cognitive autonomy.

Chapter 6: On Hephaestus' Golden maidens and built agency.

> *"...but there moved swiftly to support their lord handmaidens wrought of gold in the semblance of living maids. In them is understanding in their hearts, and in them speech and strength, and they know cunning handiwork..."*
>
> Homer (Homer, 1924) Book 18, line 420.

6.0. Introduction: Hephaestus Golden maidens and agency.

The Greek god Hephaestus built himself two female automatons made of gold to help him with his tinkering and crafting. The two robots were said to have minds and to be filled with wisdom (Homer, 1924). Given that Hephaestus was the god of mechanical crafts, it is safe to assume that he achieved this feat through engineering rather than through his divine godly power. In which case, Hephaestus would have had to contend with and solve some of the problems we have tackled in this thesis. He would have had to find solutions for how his two golden maidens would ground their mental computational structures with meaning from the world, and solutions for how they could effect enough self-control of their mental processes to qualify being credited with their own wisdom.

Yet, even with all the skill that Hephaestus was renowned for, it is not hard to imagine a visitor to his smithy acknowledging that the golden maidens were impressive, self-powered automatons that could perform several tasks, but still reserve doubts about the nature of their intelligence. Were the maidens clever in the sense of being complex wound-up contraptions that performed programmed tasks appropriately? In which case, they are not clever in the sense that humans are clever. A lot of the wisdom they are credited with would rightly belong to Hephaestus who had managed to translate solutions to various tasks into a mechanical spool that unspooled appropriately when and as demanded by the environment. Or, where they clever in the sense that they perceived the environment and set their own goals to meet the demands it posed? In which case they were clever like humans and would justify their claim to wisdom.

The question that the visitor to Hephaestus' smithy is wrestling with is a question of whether the golden damsels could take intentional actions. Many believe that AIs are not capable of performing intentional actions (Xabier E. Barandiaran, 2009, p. 382). This argument makes the case that we cannot we or should not attribute mental states that furnish intentional actions to AI? This chapter will argue that we can, and that we should if the AI is appropriately designed or built. It will lay out this argument in three sections.

Leading up to this chapter, the thesis has made the case that a path to type of autonomy that AGI would find interesting lies in replicating two functional roles of qualitative states in cognition. Two principles, the Analog Principle (AP) and the Internal Model Principle (IMP) have been proposed and defended as principles that could be used to replicate these functional roles in a computational AGI model. The first section of this chapter, 6.1, will define a minimally viable AGI model that incorporates both the AP and the IMP. There are two types of 'can exist' arguments in philosophy. The first asks whether the phenomena described is theoretically possible, while the second asks whether it is practically possible. Up until this point, the thesis has been making a theoretical case for the principles. However, to demonstrate that the model will account for intentional agency, an argument of the latter kind would be most effective. Daniel Dennet claims that a model of an entire cognitive organism is the best evidence for a functionalist theory (Dennett, 1978). In this section, the thesis will describe the cognitive model so that we can dissect it and peer inside for intentional agency. Section 6.2. will weigh the states and functioning of the cognitive model described against philosophical arguments that define intentional actions. The section will show that there is no coherent reason we should not attribute intentional action to the model.

6.1. The AP & IMP in a Computational model.

In seeking to judge Hephaestus' Golden Damsels for intelligence that is like human intelligence, the visitors to Hephaestus' smithy are seeking to weigh the damsels for intentional actions. Given that intentional actions appear to be important for the type of general intelligence that Grand Ambition finds interesting, it is important to verify whether the IMP and AP supported computational self-sufficiency can account for intentional action.

The easiest way to meet this challenge is to define a minimally viable product (MVP) of an AGI that incorporates the AP and the IMP, and to interrogate it to see whether it satisfies accounts of intentional actions. An MVP is a concept borrowed from product R&D. It is a proto product with just enough functionality to demonstrate a key feature. In this case our interest is in the computational self-sufficiency that can account for intentional actions. The MVP described in this chapter is not a full model of an AGI.

The reason it is important to define an appropriate model to inform the evaluation is because the AP and IMP are design principles. They are only informative if they have been applied appropriately. As an example, consider the concept of an arch. An arch is a curved symmetrical structure. A well-known and widely utilized design principle in construction is that an arch can be used to support large loads. This is because gravity holds an arch in compression which ensures that arches are not exposed to tensile stress. The arch principle by itself cannot account for a complete structural form like a bridge, church, or vault, but if applied in the right way in a construction project it can be used to make sturdier, safer, and longer lasting bridges, churches, vaults etc. If, however, one placed an arch on the roof of their house, it would not be the correct application of the arch principle. Such a use of an arch would not be indicative of the veracity of the arch principle.

Likewise, to properly analyse if the AP and IMP can account for intentional actions, one must build an appropriate use-scenario of the two principles and interrogate it. The next section will set the stage by describing a scenario, and a theoretical model of an AP and IMP ameliorated computational solution. The last section will critique the act.

6.1.1. The Scenario.

An ecological engineer has discovered a sealed off real world micro-environment populated by three types of items: Triangles, Squares and Circles. The ecological engineer can observe the environment through a microscope. After observing and measuring the environmental conditions, they quickly realize that the temperature in the environment is steadily decreasing. If the steady decline in temperature is not halted, the environment will be destroyed. The ecological engineer has not satisfied their inquisitiveness for the environment. They determine to set about preserving the environment by raising its temperature. Given the size of the micro-environment, the ecological engineer cannot engage directly with the environment. Being inventive, they determine that the best way to approach the problem is to insert an intelligent micro-bot that can navigate and engage with the environment in a manner that would have the desired effect of raising the temperature in the environment. The suitable bot must be small. Once deployed, it is beyond the reach of the ecological engineer's intervention. So, it must be both self-regulating and self-maintaining while it goes about the business of raising the temperature in the environment. In other words, they want their bot to make decisions for itself that will help it raise the temperature in the environment and that will sustain and maintain itself in the environment. The ecological engineer aims to build an intelligent bot that will work independently in the specialized environment.

6.1.2. The Environment.

The specialized environment in which the intelligent bot is to be deployed is dynamic. The three types of items that populate the environment, Triangles, Squares and Circles, pop-up and move about the environment continuously. The emergence and movement of the items are random except for one predictable feature which is that by its nature the environment produces mostly Triangles.

Each item carries unique characteristics determined by their shape.

- **The Triangles** have a negative heat signature and carry a negative electric charge.
- **The Squares** have a low to mild heat signature and carry a mild positive electric charge.
- **The Circles** have an extremely high electric charge and a high heat signature. As a result of these two characteristics, they cause damage on contact with other bodies.

These characteristics are proportional to the size of the item. For example, larger Triangles are colder and carry less electric charge than smaller Triangles, larger Squares are warmer and carry more electric charge than smaller Squares. The larger Circles are hotter and do more heat damage than smaller Circles.

The items also behave in predictable ways on contact. They transfer some of their electric charge onto the object that makes contact. After transferring their electric charge on contact:

- **The Triangles** are annihilated and disappear.
- **The Squares** lose their positive charge and turn into Triangles.
- **The Circles** are downgraded into Squares.

Table 6.1. Characteristics of the micro-environment.

ITEM	Thermal signature	Energy characteristics	Post-contact.
Triangle	Cold (x < 0° C)	Negative electric charge.	Removed from environment
Square	Mild (0°C < x < 50°C)	Mild electric charge.	Demoted into triangles
Circle	High (x > 50°C)	High electric charge	Demoted into squares

6.1.3. Affordances in the Environment.

An affordance is a quality, characteristic, or property that an environment offers a subject who interacts with the environment to inform its needs or ends. The term was coined by James J. Gibson in The Senses Considered as a Perceptual System (Gibson, 1966). In a later publication, Gibson defined what he had in mind when coining the phrase.

> *"The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that*

> *no existing term does. It implies the complementarity of the animal and the environment* (Gibson, 1979, p. 127)."

There are two things worth noting about Gibson's use of the word affordance. The first is that the term is not limited to positive characteristics in the environment. Gibson states that affordances in the environment might be furnished "*either for good or ill* (pg.127)*"*.

The second is that an affordance is defined from the perspective of the subject interacting with the environment. Gibson states that he intends his use of the word to define *"a complementarity of the animal and the environment* (p.127)*."* Gibson is stating that an affordance is such only if it is a property of the environment that furnishes the subject's interests. In more succinct terms, a characteristic or property of an environment is an affordance if it is useful and useable from the perspective of the subject who interacts with the environment.

Does the micro-environment offer any characteristics or properties that can be complementary to a subject, in this case an intelligent bot, who has been motivated to increase the temperature in the environment?

Some obvious affordances stand out.

The Triangles in the environment carry negative thermal properties that are primarily responsible for bringing down the aggregate temperature in the micro-environment. The proclivity of the environment in producing Triangles account for why the temperature in the environment is trending downwards. The most efficient way of raising the temperature in the environment is to remove as many of the Triangles as possible. Conveniently, contact with a Triangle annihilates it from the environment. Therefore, for a subject that is sufficiently motivated to seek a rise in the temperature, the Triangles furnish an affordance. Contact with a Triangle will remove it from the environment and raise the aggregate temperature.

Frequent contact with a Triangle raises a secondary problem. The Triangles carry a negative electric charge, and they transfer this negative charge on contact. This means that the more Triangles a subject consumes, the more energy is drained from the subject. This is a problem because a subject, whether it be an artefact or an organism, requires energy to draw from to function. If depleted, as would be the eventual outcome of too much contact with Triangles, the subject will be immobilized. This problem is compounded by the fact that the ecological engineer cannot intervene with the subject once it has been deployed in the environment. The subject must therefore be equipped to seek out an affordance to solve this secondary problem through its behaviour.

Once again, the environment offers a potential affordance for the subject that can help solve this secondary problem. The Squares carry a mild positive electric charge that is transferable on contact. However, the subject must self-regulate its charging needs. Too much contact with the Squares might cause it to over-charge which would damage it. Secondly, contact with a Square will change the Square into a Triangle which has the effect of lowering the aggregate temperature in the environment. This is counterproductive to raising the temperature. Seeking out contact with Squares in the environment therefore represents an affordance that is potentially both positive and negative. The subject interacting with the environment must be able to self-regulate its engagement with the affordance, by only seeking out contact with Squares when necessary and in proportion to its energy needs. In this way the subject will avoid damage from over-charging and manage the cooling effect that the topping up process has on the environment.

Lastly, as part of its self-maintaining mandate the subject must avoid as much damage as possible. The costliest form of damage will be occasioned by contact with Circles. The larger the Circle, the more the damage suffered by the subject. Damage will markedly reduce the efficiency of the bot by immobilizing it until it has recovered. The recovery period is proportional to the damage suffered. The more severe the damage the longer the immobility. A critical amount of damage can be fatal – permanent immobility. Since the environment is more prolific at producing Triangles, any reduction in efficiency on the part of the bot will mean Triangles in the environment increase and the aggregate temperature in the environment will continue to drop. Haphazard contact with Circles is therefore potentially expensive for the health of the bot and for the environment.

Table 6.2. Affordances in the Micro-environment.

The micro-environment furnishes at least three items that offer affordances that will be both useful to and useable by the bot.

ITEM	Affordances
Triangle	Positive: Capturing reduces temperature in the environment. Negative: Drains energy from the subject.
Square	Positive: Contact recharges energy level. Negative: Contact will reduce temperature in the environment.
Circle	Positive: Avoiding contact maintains health. Negative: Action to avoid contact distracts from primary motivation.

Armed with knowledge of affordances in the environment, the ecological engineer sets out to build an intelligent artefact that can be deployed in the micro-environment to raise the temperature. The ecological engineer christens the project and the bot they aim to build the **Autonomous Geo-Embedded Nano Thermostat** bot or **AGENT-bot** in short. The AGENT-bot is a simple micro-artefact that has basic motion capability that allows it to stay in place, approach, or retreat in any direction. The ecological engineer aims to equip AGENT-bot with the intelligence to make the right choices, either stay, approach or retreat, in a manner that will yield a rise in the temperature of the micro-environment.

6.1.4. AGENT-bot's Umwelt.

Of primary interest to the ecological engineer should be to arm the AGENT-bot with access to the affordances. To make use of the affordances in the environment, the bot must first perceive them. Kant identified two faculties in cognition, the faculty of sensibility and the faculties of understanding and reason (Kant, 1781 [1998]). The faculty of sensibility allows a cognitive agent to perceive (sense) items in its environment. It is based on the faculty of sensibility, that a cognitive agent can build understanding and knowledge. The first task at hand for the ecological engineer should be to equip AGENT-bot with a faculty of sensibility. In the interest of economy, it would be best to focus the faculties of sensibility on the affordances in the environment.

What the ecological engineer is trying to do by equipping AGENT-bot with a faculty of sensibility that prioritizes affordances in the environment is to establish an Umwelt for the bot. The term 'Umwelt' describes the internal appearance of the world to an organism. Organisms will have different Umwelten even though they share the same environment. The organism's Umwelt is determined by its relation to its environment. It is a sensitivity to the elements in the environment that are *"carriers of significance"* which interest the organism (Agamben, 2003, p. 46). As an example, Jakob Johann von Uexküll speculated that the Umwelt of a tick consists of at least three carriers of significance. The odour of butyric acid from mammalian pores which alerts the tick that a prey is within reach and prompts it to fall onto the prey. The temperature of 37°C which indicate to the tick that the heat source is a warm-blooded mammal and prompts it to seek out soft cutaneous tissue to start feeding. The hairy topography of the mammal that guide the tick in navigating to a feeding spot (Agamben, 2003, p. 46).

Uexküll was describing the Umwelt long before James J. Gibson coined the term 'affordances,' however it is clear that the term 'carriers of significance' relates in a meaningful way to Gibson's term 'affordances.' The term 'carriers of significance' references a marker that identifies or indicates an affordance. The odour of butyric acid, the 37°C temperature, the feel of mammalian hair are all sensory markers, the perception of which indicate to the tick that there are affordances within proximity. Without a faculty of sensibility on these markers, the tick could not function as an agent in its environment.

Beyond simply identifying the affordances in the environment, the Umwelt serves to create an internal model of the environment that informs decision making. An Umwelt is a subjective model of the world (Kull, 2010). It is the world as it is represented by the faculties of the organism. The Umwelt maps the exploitable features of the environment that are perceptible by the organism. In mapping the world, the Umwelt equips the organism with underived meaning by indicating the meaningful features of the world, as well as their location in relation to the organism. This allows the organism to make critical decisions.

Take for example Uexküll's tick. Its Umwelt consist of a sensitivity to butyric acid. We can only speculate how this sensitivity is represented in the tick's Umwelt, but we can be certain it informs the tick by way of indication that there is food within reach. It also indicates in which direction in relation to the tick the meal is located. These are both important pieces of information that determine how the tick acts, e.g., in which direction the tick moves to access the meal. In much the same way, the temperature of the mammal and the way the mammalian hair grows are certainly represented in the tick's Umwelt and the representation of them in the Umwelt serve to inform both meaning and action for the tick. The Umwelt not only indicates notable features of the environment, but it also maps the environment for the organism to inform decision making.

The ecological engineer must equip AGENT-bot with an Umwelt. AGENT-bot's Umwelt must furnish access to the environment in the form of a perspective that indicates the affordances in the environment and maps them in relation to the bot. The Umwelt should have two defining characteristics.

- A sensory model that indicates meaningful aspects of the world.
- A sensory model that maps the environment in reference to the experiencer and thereby informs decision making.

The AP and the IMP can be used to perform both functional roles.

6.1.5. The AP as a sensory model that informs meaning from the world.

The Analog Principle explains how computational items can acquire underived meaning from the world. It states that to automate this grounding process, an analog mechanism is essential. This mechanism uses an analog model to represent discernible, differentiable, continuous and variable signals from the world, which serve as the grounds of meaning. The model then maps this meaning onto computational items.

An analog mechanism has two main facets: a model with qualitative properties (such as discernibility, differentiability, continuity, and variability) and computational features that are synchronized in a meaningful way.

Consider examples like a clock and a thermometer. In a clock, the angular movement of the hands (qualitative *ad litteram*) is synchronized with the numerical markings on the clock face in a meaningful way. Similarly, in a thermometer, the expansion and contraction of mercury (qualitative *ad litteram*) align with the temperature scale on the tube.

The analog mechanism is straightforward. Its qualitative properties are designed to analogously represent the continuous signal from the world that imparts meaning. By capturing the essential features of the worldly signals, the qualitative properties transfer meaning to the computational facet of the model, which is synchronized with the qualitative aspect.

For example, the movement of a clock hands is set to model the sun's movement around a sundial, thereby transferring meaning to the numbers on the clock face. Likewise, the mercury's contraction and expansion represent temperature changes, transferring meaning to the numerical markings on the thermometer.

The Analog Principle can be used to equip AGENT-bot with a sensory model of the world that highlights computationally meaningful aspects of the world, and transfers this meaning onto computational items that process behaviour. This would entail that the ecological engineer equips AGENT-bot with an analog mechanism that can perform three steps.

- **Sense** the relevant continuous signals from the environment.
- **Model** these continuous signals with qualitative properties.
- **Transfer** meaning from these qualitative features onto synchronized computational items.

In AGENT-bot's environment, the key elements offering affordances are Circles, Triangles, and Squares. Each are distinguished by unique discernible and differentiable temperature signals. These distinct signals can be utilized in the analog mechanism.

To clarify the three-step process, the following section will discuss specific design choices an ecological engineer might make to fulfill these steps.

1. Sense.

To track the affordances in the environment, AGENT-bot must be set up to sense the thermic properties from its environment. This means it must be able to detect the heat signals from the items in the environment. To sense the thermic properties, AGENT-bot can be equipped with heat sensors that scan the environment for thermic properties.

2. Model.

AGENT-bot must be able to model the thermic properties from the micro-environment in some form that ensures analogous fidelity to the continuous signals from the world. For example, AGENT-bot could be equipped to model the qualitative properties of heat with qualitative properties of light. The light signals must mirror the heat signals in a meaningful way. E.g., the hotter the signal, the brighter its indicative light signal. In this manner the lights are an analogous model of the world.[59]

3. Transfer meaning.

The qualitative model in step (2) must be meaningfully synched to computational items. For instance, in this scenario, computational items could measure luminance. AGENT-bot could use photometers to assign luminance values to features represented in the light-based model. These luminance values could then be processed by an algorithm that groups and classifies them into distinct types. For example, items with a luminance value corresponding to real-life temperatures below 0°C would be classified differently from those with values representing temperatures between 0°C and 50°C.

Machine learning algorithms trained through unsupervised learning are particularly effective and reliable for classification tasks like this. The classified items gain meaning because they represent actual features of the world. This is an important point. The meaning attached to these computational items references real-world features (such as properties of worldly items) rather than meaning in a derived sense or meaning that relates to language.

Through the Analog Principle, AGENT-bot can self-ground its computational items with direct, causally imparted meaning from the world. This meaning directly references real-world features, allowing AGENT-bot to sense its environment and extract meaningful information that guides its internal computations and behaviour.

6.1.6. The IMP as a sensory model for enabling decision-making.

The Internal Model Principle as applied in this thesis is faithful to the Internal Model Principle in control theory. The Internal Model Principle in control theory states that to engineer for self-

[59] It's important to note that qualitative properties other than those of light can also be used to model signals from the world. The key requirement is that the chosen properties maintain an analogous fidelity with the signal. This paper uses light as an example because it's likely easier for both writer and reader to conceptualize a relationship between light and heat. However, other qualitative properties for signals such as sound, pressure, or vibrations, could work just as effectively. The Analog Principle is agnostic about form.

control in a system an engineer must build into the system itself an internal model of the dynamics that generate the signals that influence or motivate control instructions (Wonham, 2018). This internal model is a controller that delineates the system itself, depicts, and monitors internal and external changes that affect the system to inform self-control decisions (Camacho & Bourdons, 2004).

The line put forth and defended in this thesis is that the Internal Model Principle informs the functioning of self-control in cognition. The mind features an emergent mental entity called the inner self. While the ontic nature of the inner self is the subject of debate, there is some agreement that its defining characteristic is that it is the entity that should be credited with aspects of control in cognition. This consensus is evident in terms such as self-control, self-determination, self-propelled. In all of which the prefix 'self' indicates that the control, determination, or propulsion being described is creditable to an internal entity called the self.

In cognition, the minimal construal of a self consists of two definitive features. These are the sense of body ownership and a sense of agency (Gallagher, 2000). The sense of body ownership are feelings that relate to the demarcation of the body, situatedness in an environment, and perspectivalness. The sense of agency is a feeling of being able to initiate and control actions through exercising decision making. The ability to be able to discern and differentiate is essential to enacting motivated choices.

Together, the sense of body ownership and the atomistic conditions for the sense of agency constitute a minimal self that serves the functional roles of creating a delineation of the body, an awareness of environmental dynamics, a perspective that centralizes both interoceptive and exteroceptive information, an ability to discern and differentiate information, and the grounds on which to exercise motivated choices based on those discerned differences. These functional roles compare favourably to the role of an internal model in control theory.

Qualitative experiencing and the Umwelt.

The description of what the self affords to cognition is not only in conformity with the functional roles of a controller in the IMP, but it also rhymes with a function of an Umwelt. The self emerges from qualitative experiencing. The very act of having a qualitative experience yields a sense of body ownership and an emergent sense of agency.

As a reminder, one of two defining characteristic of the Umwelt is that it is a sensory model of the world that maps the environment in reference to the experiencer. The model serves to orient the organism and inform decision making. To support this line of argument, consider Uexküll's tick. It uses its sense for butyric acid to orient itself in the environment and make decisions that inform what the tick should do in response to the experience. One can build on this example to highlight how Uexküll's tick's experience of butyric acid furnishes it with a (1) sense of body ownership and a (2) atomistic sense of agency.

(1) Sensory experiencing and the sense of body ownership.

On having an experience related to butyric acid, Uexküll's tick will contend with three consequent and emergent implications.

i. Demarcation of body.

Uexküll's tick will have some form of qualitative experience as a marker of its experience of butyric acid. As Nagel argued in his 'What is it like to be a bat (Nagel, 1974)' we cannot know how this experience will feel for the tick. However, one can abduce that the experience will present as an experience belonging to the tick. Drawing from human experience, bodily sensations demarcate a part or parts that hosts the sensation and mark them off as belonging to the experiencer. A qualitative experience means that the subject of the experience recognizes the experience as being borne on a body that is accessible to the experiencer and that is separable from the rest of the world.

ii. Situatedness.

Secondly, Uexküll's tick will have a sense of being situated in an environment. Being able to demarcate the body as the bearer of experience implies that the demarcated body is accessible and recognizable as a thing distinct from the surrounding environment. The surrounding environment is presented as foreign to the demarcated body. However, while the environment is presented as foreign to the demarcated body, it is nevertheless recognizable as an outside source or cause of some of the signals that yield experiential information. The tick's sensory organs are gateways for worldly information from the environment. The information introduced through the sensory organs, in this case the presence of butyric acid, is represented in the form of sensory information. This creates a sense of the demarcated body's situatedness in an environment that feeds the body with sensory information.

iii. Perspectivalness.

Thirdly, the experience will imply that Uexküll's tick will orient itself in the environment as a centre of perspective. The previous two points have made it clear that sensory experiencing implies the demarcation of body, which marks the body of from the environment in which it is situated, but also makes it apparent that the environment can feed the demarcated body with sensory information. The ability to become aware of sensations from the environment is called perception. The centre of perspective is the loci of experiencing sensory information and therefore the loci of perceiving the environment that cause the sensory information. The ticks sensing of butyric acid will present as a first-hand perspective on the causes of sensory information. This first-hand perspective is useful for helping the tick orient itself in its environment.

The experience of sensing butyric acid has therefore armed the tick with demarcation of body, situatedness and perspectivalness. Taken together they form a sense of body ownership.

(2) Sensory experiencing and the Sense of Agency.

The sense of agency is emergent from two implications of having a sensory experience. On having a sensory experience related to butyric acid, Uexküll's tick will contend with two implications.

i. Discernment.

The first implication is that Uexküll's tick will be able to discern the experience. Holding any form of sensory experience implies the ability for the experiencer to discern the experience. A sensory experience is made discernible by the qualitative property of discernability. The immediate implication of a qualitative experience is that the experiencer can discern between having an experience and not having an experience. In addition, Uexküll's tick will also be able to discern between different types of experiences. This is because qualitative experiences also have the property of differentiability. Uexküll's tick's experience of butyric acid will be different from its experiencing of the 37°C which relates to mammalian body temperature. In cases where the experiencer is capable of hosting more than one type of experience, the implication is that the experiencer will be able to discern one experience from another.

ii. Conditions to exercise choice.

Uexküll's tick's ability to discern between experiences imply it has the necessary conditions to exercise motivated choice. There will be more said about the operationalising of exercised choice below, at this point it is important to highlight the fundamental condition on which exercising choice is based.

As stated in the previous chapter, this point is most clearly understood from its converse position. In the absence of discernible differences, there is no basis on which to exercise choice. For example, in conditions of perfect uniformity, in cases where there are no differentiable or distinguishable features, there are no grounds on which to exercise choice. The conditions for exercising choice are dependent on there being discernible and distinguishable differences between the items subjected for choice.

If Uexküll's tick could not discern between a sensory experience for butyric acid, and for example having no experience, then it would not have the conditions in place to respond to butyric acid. Likewise, if it could not differentiate between a sense for butyric acid or a sense that indicated the presence of a natural predator, the tick would not be able to choose to respond to the beneficial signal and avoid the harmful signal. The latter example leads nicely to the matter of operationalizing motivation.

To motivate choices, cognitive agents use a different type of experiencing. Cognitive agents use hedonic reward signals to motivate choices between perceived stimuli. Hedonic reward signals refer to the sensations of pleasure or discomfort directed towards maximizing exposure to favourable stimuli while minimizing exposure to unfavourable stimuli (Schultz, 2015). Hedonic reward signals are the operative mechanism for associative learning (Kolb & Whishaw, 2001, pp. 439-9). Associative learning is the process through which organism acquire a knowledge of relationships between stimuli and events in their environment (Christian, 2010). It is the primary process that cognitive agents use to adapt existing behaviour or develop new behaviour. In this process, the cognitive agent adjusts its behaviour based on whether the formed association has either positive or negative rewards (Christian, 2010).

Uexküll's tick knows that approaching a signal from the world that indicates butyric acid leads to a positive reward, in this case food and its accompanying feelings of satiation. Uexküll did not mention a natural predator for his tick, but quick research reveals that ants are a natural predator for most ticks. Most ants use formic acid both as a venom and pheromone. We might therefore imagine that Uexküll's tick might learn through associative learning that approaching a signal from the world that indicates formic acid represents a negative reward, in this case danger and its accompanying negative feelings. The tick is therefore able to make behavioural choices based on motivations influenced by positive or negative rewards.

However, the ability to make these choices are fundamentally dependent on the ability to discern and differentiate between incoming stimuli. If the tick had no way of discerning and differentiating between experiences arising from a signal for butyric acid and a signal for formic acid, it would not be able to exercise a choice between the two experiences.

Uexküll's tick's experience of butyric acid has armed it with the fundamental conditions for a sense of agency. Its ability to have an experience of butyric acid implies that the tick can discern and differentiate between experiences, and that it has created the grounds on which it can make motivated decisions.

In the example above it is demonstrable, that for Uexküll's tick, possessing an Umwelt implies the emergence of what is recognizable as an emergent inner self. The emergent inner self is what is credited with effecting Uexküll's tick's decision making. It serves an analogous role to an internal model in that it demarcates the body and centralises and integrates information from the demarcated body and the environment that Uexküll's tick's leverages to makes decisions.

6.2.7. Building AGENT-bots Umwelt

Given that the functional implications of experiencing (having an Umwelt) are comparable to the functional role of an internal model in control theory, the ecological engineer can turn to the IMP to build AGENT-bot's Umwelt.

The task at hand is to interpret the implications of qualitative experiencing in terms of replicable functional roles that can be realized by applying the IMP. Qualitative experiencing yields the sense of body ownership and the sense of agency. The sense of body ownership gives rise to a delineation of the body, an awareness of environmental dynamics, and perspective that centralizes both interoceptive and exteroceptive information. The atomistic sense of agency gives rise to an ability to discern and differentiate information, and the grounds on which to exercise motivated choices based on those discerned differences. All of which are functional roles that can be replicated with an internal model in the IMP.

The functional roles that the ecological engineer is looking to replicate in AGENT-bot are derived from the implications of qualitative experiencing and are as follows:

Table 6.3: Functional roles of a sense of Body ownership.

Implication of qualitative experiencing	Functional role in realizing sense of body ownership
Demarcation of Body	Create a delineation of the body and track interoceptive information.
Situatedness	Create awareness of environment and track shifting dynamics from environment
Perspectivalness	Centralize interoceptive and exteroceptive information and create awareness of body position in environment.

Table 6.4: Functional roles of a sense of agency.

Implication of qualitative experiencing	Functional role in realizing sense of agency
Discernability & differentiability	Ability to discern and differentiate between different signals.

Conditions for exercising choice.	Create the foundations for exercising choice
Hedonic motivations	Exercise motivated choice

The Internal Model Principle (IMP) states that, for a system to achieve self-control, it must include an internal model of the dynamics that generate the signals driving its control instructions (Wonham, 2018). The control system in the shape of an internal model delineates the system itself, integrates information about external factors and internal states, then issues control instructions based on this information. In more advanced systems, the internal model also monitors feedback on its instructions, enabling continuous error correction.

Conveniently for the ecological engineer, in opting to utilize the Analog Principle they have already made choices in their design for AGENT-bot that will equip it with an internal sensory model. AGENT-bot is equipped with access to modelled light signals that represent the qualitative properties of heat from its environment with qualitative properties of light. The light signals that make up the internal sensory model are an analogous representation of the heat signals from the world. This internal sensory model can be co-opted to perform the functional roles of experiencing in effecting self-control.

1. Demarcation of body.

To create a delineation of the body and track interoceptive information, the ecological-engineer must ensure that AGENT-bot's internal model includes a model of the AGENT-bot itself. Since AGENT-bot's model uses luminescence to represent environmental heat properties, this self-model must also rely on distinct luminescent signals to accurately differentiate AGENT-bot from its surroundings.

The purpose of modelling environmental information is to transfer meaning from the world onto computational items that AGENT-bot uses to generate behavioural output. To process this output effectively, AGENT-bot must track both external (exteroceptive) and internal (interoceptive) dynamics. External signals are represented by steady beams, where light intensity corresponds to temperature. To distinguish its self-model from these external signals, AGENT-bot's internal model might represent itself using a distinct type of light, perhaps pulsating signals, which contrast with the continuous beams used for environmental signals. This difference in signal type helps demarcate AGENT-bot's "self" from environmental inputs within the model.

It should be kept in mind that the modelled signals serve a purpose, which is to ground information onto computational structures that compute output. The model of self in the internal model is equally pragmatic. The modelled self is intended to ground interoceptive information about the system itself into computational structures that will be critical to inform processing of the right output.

For dynamic interaction, AGENT-bot needs to track various internal states to adjust its behaviour effectively. Identifying three core states will be useful for AGENT-bot to navigate its environment: an optimal state, a low-energy state, and a damaged state. Awareness of these internal changes informs AGENT-bot's response, allowing it to adapt based on its condition.

The engineer can represent these states through distinct pulsating light patterns. For example:

- A slow, steady pulse might signal an optimal state.
- An alternating dot-dash pattern could indicate low energy, with alternation speed indicating severity.
- A rapid, steady pulse could signify damage, with pulse speed reflecting its intensity.

These modifications ensure that AGENT-bot's internal model includes a clear representation of itself, distinguishing its own states from external signals and representing each critical internal state distinctly.

2. Situatedness.

To create environmental awareness and track dynamic changes, AGENT-bot is designed to detect the qualitative properties of thermal signals and represent them as light properties within a sensory model. This light-based sensory model mirrors the environment, allowing AGENT-bot to monitor shifting dynamics accurately. As discussed earlier, this environmental model also includes a representation of AGENT-bot itself, positioned within the sensory model to reflect its real-world location. This internal model establishes a sense of situatedness, positioning AGENT-bot not only as a receiver of environmental stimuli but also as an integrated part of the environment..

3. Perspectivalness.

AGENT-bot's light-based sensory model creates a perspective on both its interoceptive and exteroceptive signals.

The model tracks interoceptive signals to give AGENT-bot a view of its changing internal states, which is essential for grounding information that informs its approach to the environment. For instance, if AGENT-bot's internal signals show it is low on energy, this information helps guide its behaviour toward prioritizing energy-gaining actions and avoiding those that might deplete its reserves. Conversely, when its internal signals indicate a healthy state (e.g., high energy, no damage), AGENT-bot's behaviour can focus on achieving its primary goal–raising the temperature in the micro-environment.

The sensory model also monitors exteroceptive signals, giving AGENT-bot a perspective on its surroundings and the environmental features that cause the stimuli. This orientation enables AGENT-bot to recognize available affordances and their spatial position in relation to itself. Through this perspective AGENT-bot has a sense of the affordances and features in the environment and can orient itself in relation to these affordances and features.

4. Discernment

AGENT-bot's light-based sensory model represents both external (exteroceptive) and internal (interoceptive) information using the qualitative properties of light. External signals are represented by continuous beams, where luminosity corresponds to the temperature of features in the environment. Internal states are represented by unique pulsating patterns, each corresponding to one of AGENT-bot's three primary physical states. These qualitative properties are crucial for enabling discernment, as they allow AGENT-bot to distinguish signals.

By modelling external signals with internal representations that faithfully mirror relational properties of real-world features, AGENT-bot can ground computational structures in a way that relates meaningfully to environmental features and their spatial relation to itself. Without discernible qualities, AGENT-bot would be unable to recognize environmental features. For example, if Triangles in the environment had no distinct thermal properties, AGENT-bot would be unable to represent them in its internal model, meaning it couldn't access information about Triangles in its computations.

Similarly, the qualitative properties of internal signals enable AGENT-bot to differentiate its internal states. AGENT-bot perceives its states as distinct pulsating light patterns, providing a unique perspective on each. Without these properties, AGENT-bot would lack the ability to identify and computationally process each state as a distinct, useful condition.

5. Differentiability.

AGENT-bot's internal sensory model not only allows it to detect both internal (interoceptive) and external (exteroceptive) signals but also to differentiate between them. This differentiability is achieved through qualitative distinctions in the model. For instance, AGENT-bot can distinguish and categorize Triangles, Squares, and Circles based on the unique ways these shapes are represented in its internal model. These internal distinctions faithfully correspond to the actual differences in features found in the environment.

The ability to discern and differentiate between signals is essential, as it creates the conditions needed for making choices. Making a choice involves selecting between two or more options, each of which must be distinct to allow for comparison. In AGENT-bot's model, choices are based on individuated features made visible by distinct light properties. With its ability to distinguish between Triangles, Squares, and Circles, AGENT-bot has the foundational conditions required for processing motivated choices among these three items.

6. Motivated choice

The design choices made by the ecological engineer have given AGENT-bot an internal sensory model that enables it to distinguish between different environmental signals. This model provides AGENT-bot with a perspective on meaningful items in its surroundings. The next step is to equip AGENT-bot with the motivation to make choices within its environment.

Inspired by the hedonic reward system in cognition, the ecological engineer can implement motivated decision-making in AGENT-bot through reinforcement learning. In cognitive agents, hedonic rewards drive associative learning by guiding choices in response to the environment.

Reinforcement learning (RL) functions similarly to hedonic rewards in cognition. It is a training method that supports associative learning, where the RL agent learns to make behavioural choices based on environmental feedback in the form of reward or punishment signals. The RL agent is designed to maximize cumulative rewards from its environment. In this setup, reward signals are tied to desired goals, while punishment signals indicate obstacles to those goals. By interacting with its environment, the RL agent learns to make choices that optimize cumulative rewards.

The ecological engineer wants AGENT-bot to be principally motivated to increase the temperature in the environment. To this end, the engineer might pair a primary reward signal to the goal of increasing the temperature in the environment. To maximize the reward signal, AGENT-bot must learn to make choices in its environment that cumulatively yield a temperature spike.

We should be reminded that AGENT-bot has basic motor capabilities, it can stay, approach or retreat. It has also been designed with an internal model that presents a perspective on its environment that allows it to discern and differentiate the items in its environment. Through reinforcement learning training AGENT-bot can learn associatively whether to stay, approach, or retreat from the items in the environment in a manner that will ensure it attains cumulative rewards that result from the end goal of increasing the temperature in the environment.

An example of a choice that AGENT-bot can learn in this case might be a choice to approach the Triangles and through contact annihilate them from the environment. Note the speculative tone I have assumed in this example. This is because this approach is dependent on associative learning. The ecological engineer does not prescribe AGENT-bot's actions. They are learned from the feedback in the environment. It is quite possible that AGENT-bot might find actions that are more optimized to raise the temperature in the environment than approaching Triangles. With this method, AGENT-bot's learned instrumental behaviour is creditable to AGENT-bot.

Besides raising temperature, AGENT-bot has additional objectives: maintaining its energy levels and protecting its health. These goals can also be motivated through reinforcement learning by pairing reward and punishment signals with the appropriate objectives. For instance, if punishment signals are tied to declining energy or health, AGENT-bot should learn behaviours to replenish or preserve these resources.

Reinforcement learning also supports AGENT-bot in prioritizing its goals based on its current physical state. If AGENT-bot's health and energy are at optimal levels, it should focus on increasing the temperature. When energy is low, it should shift to seeking energy sources. If damaged, it should prioritize recuperation. AGENT-bot's internal sensory model provides interoceptive view on its states, such as health and energy, through distinct pulsating light patterns. By processing these light patterns, AGENT-bot can learn to adjust behaviour dynamically based on internal feedback.

The qualitative properties of the pulsating light patterns, enable AGENT-bot to distinguish between its internal states. AGENT-bot has *ad litteram* qualitative experiences that inform its attitude to its environment. Also note that meaning of these internal states are not prescribed to AGENT-bot. They are learned through associative learning from feedback in the environment. AGENT-bot learns to associate its attitudes towards the environment to its internal states by the rewards and punishment signals it receives from interacting with the environment.

In this manner, by way of a light-based internal model, the ecological engineer can equip AGENT-bot with an Umwelt. An Umwelt is a subjective model of the world that is represented by the faculties of the experiencer. AGENT-bot's faculties represent the world in terms of properties of light. This is different from organic faculties that represent the world in terms of phenomenal properties or qualia. However, while AGENT-bot's internal model is materially and experientially different, there is a functional symmetry with organic Umwelts. Both are a sensory based models that ground computationally meaningful information from the world and map the environment in reference to the experiencer, and thereby inform the experiencer's decision making. AGENT-bot senses, gains meaning, and processes information from its native environment in terms of properties of light represented in its internal model.[61]

6.1.8. Is AGENT-bot a minimal example of an AGI?

Is AGENT-Bot an AGI? An immediate criticism of AGENT-Bot could be that it seems to have generalization that is like multi-game playing agents such as MuZero (Schrittwieser, et al., 2020). A criticism that was labelled on these AI agents earlier in the thesis was that they are equipped with one solution that has broad applicability in a narrow domain. These agents have a non-transferable competence that is limited to generating action policy in a 2D pixelated virtual environment that returns rewards according to game rules. This generalization is different form the adaptable, flexible, and transferable generality that is sought in the Grand Ambition's definition of general intelligence. The temptation is to view AGENT-Bot in the same lens.

Two things can be said in defence of AGENT-bot.

The first is that AGENT-bot should not be thought off as operating in a virtual environment. The thought experiment asks you to think of AGENT-Bot as a real nanobot with simple motor capabilities that allow it to stay in place, approach, or retreat in any direction. While AGENT-bot is optimized for its environment, it could very well be placed in another micro-environment, even one more complex, and transfer its learnings to that environment with varying and dependent degree of success.

Secondly, it should be kept in mind that AGENT-bot is intended as a minimum viable product (MVP) of an AGI. An MVP in product design and development is a proto product with only the essential features needed to test a core hypothesis. Here, the focus of describing AGENT-bot was to demonstrate the computational self-sufficiency that would be compelling for AGI

[61] It is important to stress that equipping AGENT-bot with a light-based sensory model is a design choice that the author has imparted on the ecological engineer. An internal model based on some other qualitative properties would work just as well if it could faithfully model the salient features of the environment. Light is an easy property to juxtapose against heat, which is the indicator of interest in the environment. The danger in describing a light-based sensory model is that from human experience one tends to associate light with the capacity of seeing. This might induce the reader to think that the purpose of the internal model is a visual display for something, or someone, and thus expose the proposal to accusations of the homunculus fallacy. To ward off this accusation, it is important to stress that the purpose of the internal model is NOT to display. The qualitative properties of light in the model serve the functional role of grounding meaning and supporting the emergence of the functional roles necessary for effecting self-control. Both these functional roles could be realized by way of a sensory internal model built on some other signal with qualitative properties.

research. The reader should not think of AGENT-bot as having solves all the problems related to AGI. A degree of charity is warranted for its shortcomings.

However, with that said, for us to properly weigh if AGENT-Bot's computational self-sufficiency is of a type that will interest AGI, it has to be shown to support some aspects of general intelligence, specifically the ability to take intentional actions.

The ecological engineer has designed an artefact that is computationally self-sufficient in a way that seems relevant to AGI. AGENT-bot has at least three broad objectives. To take instrumental actions that will raise the temperature in the environment, manage its energy needs, and maintain its health. In pursuing its goals, it confronts problems in forms that are framed by the environment and its internal state. These problems are dynamic, and in response to them, AGENT-bot must match the dynamic framing of the problems in the environment with dynamism in problem solving efficacy.

Secondly, AGENT-bot appears able to learn to solve unfamiliar problems. All of AGENT-bot's problem-solving is learned rather than pre-programmed. The ecological engineer can only hypothesize how AGENT-bot might tackle challenges in its environment. Its actual solutions emerge from interaction and feedback rather than being predetermined.

For example, while it may seem intuitive for AGENT-bot to approach and neutralize Triangles to reduce temperature, AGENT-bot might discover more effective strategies through experience. AGENT-bot's problem-solving process depends on its internal state, how it perceives the environment, and the feedback it receives. This flexibility allows AGENT-bot to potentially encounter new problems that were not anticipated by the engineer.

As an example, imagine AGENT-bot is simultaneously damaged and low on energy. Each of these states has a unique sensory representation (unique light patterns). Combined, they may form a new, distinct sensory pattern, creating a novel experience for AGENT-bot. How will it respond to this new situation? The ecological engineer can only speculate. AGENT-bot might approach a small Triangle near a large Square to use the Triangle to cool itself with minimal energy loss, and the Square to replenish energy. Or it might collide twice with a large Square: first to gain energy, and then, after it transforms into a Triangle, to soothe its damage. AGENT-bot's actions in response to this new state will ultimately be guided by the feedback it receives and its learned responses over time.

Another argument for AGENT-bot's potential for general intelligence is that its human handlers must rely on reasoned speculation to predict its actions. Take for example the expectation that AGENT-bot's go-to move will be to annihilate Triangles. This expected move is speculative, not prescribed. This may be a meaningful indicator of general intelligence, because when the handlers speculate about AGENT-bot's behaviour, they are effectively placing themselves in its position. They are imagining how they might solve the problem with AGENT-bot's abilities and under the same conditions.

For example, the thought process might go as follows: *Given that I had abilities A, was in internal state B, and the environment was in state C, my response would be action z.* Here, z represents the handler's reasoned solution to the combination of conditions A, B, and C.

This approach to predicting AGENT-bot's actions is significant in the AGI discussion because a primary goal of AGI is to achieve intelligence functionally comparable to human intelligence. If the handlers find themselves predicting AGENT-bot's behaviour by imagining how they would respond to the same challenges, it implies they recognize AGENT-bot's problem-solving capabilities as functionally similar to human intelligence.

Yet, there are some who might agree that AGENT-bot is a very impressive self-sufficient, multi-problem-solving, computational model, and yet still question whether AGENT-bot is performing intentional actions. Does AGENT-bot take intentional actions in the sense that will satisfy philosophical pondering on the matter? In philosophy, intentional actions can be and should be distinguishable from acts that are merely the result of a series of events in causal relations no matter how complex the causal chain is. The argument could be made that AGENT-bot is not capable of intentional action. In that case, AGENT-bot could not be credited with general intelligence, at least not in the form that is functionally comparable to human intelligence. The next section will critique this argument.

6.2. Does Agent-bot take intentional actions?

Given that AGENT-bot appears to have some evidence of both cross-domain intelligence and the ability to confront novel problems, the argument as to whether it's self-sufficiency qualifies it as functionally comparable to human intelligence is going to turn on whether one should credit AGENT-bot with intentional actions.

In philosophy of action, intentional agency is reserved for actors that can initiate their own actions. The dominant thought in philosophy of action is that this involves the agent owning an intention to perform the action. According to this position, whether AGENT-bot is performing intentional actions, will turn on whether we can locate ownership of a state, states, processes, or ratiocinations, which can be identified as an intention. If one cannot trace ownership of an intention to AGENT-bot, then it might be argued, that computational self-sufficiency notwithstanding, AGENT-bot is not performing intentional actions and therefore is not fit to be labelled an MVP of an AGI. It is simply a clever assemblage of causal relations that generate actions that do not fit the narrow definition of intentional actions.

Is AGENT-bot performing intentional actions? Theorizing in the philosophy of action has yielded the Standard Conception of Agency (SCA) and the Standard Theory of Agency (STA). This section will weigh AGENT-bot's behaviour against both the SCA and the STA to determine whether there is anything within both formulations that deny AGENT-bot intentional agency.

6.2.1. The Standard Conception of Agency.

In philosophy of action, intentional agency is construed as the capacity to take intentional actions. Intentional actions are defined by the Standard Conception of Action as a special form of action. The SCA typifies these narrow actions as actions that are motivated by distinct states or conditions that are labelled intentions. Intentions are related in some sense to a goal that the agent (i.e., the bearer of agency) acts towards. Intentional actions are therefore often interchangeably known as goal-directed actions. These actions are separable from activity that results from entities merely acting on each other in a causal relationship.

In the SCA framework, intention always precedes the action, and the action can be explained by referencing this intention (Schlosser, 2019). Thus, an intention should be understood as the reason behind the action or, at the very least, closely linked to a reason for acting (Mele & Moser, 1994), (Clarke, 2010), (Enç, 2003).

From the standard conception of action, one derives a conception of intentional agency. An entity has intentional agency if it has the capacity take actions that are preceded, informed, or motivated by intentions. The standard conception of agency does not commit to a particular definition of what intentions are or how they are related to reasons for acting (Schlosser, 2019). It merely aims to individuate a particular type of action. An action that is preceded and informed by a state or conditions identified as an intention, and that this intention is closely related to the reason for the action.

Does AGENT-bot conform to the SCA?

From the Standard Conception of Action (SCA), we can derive two key claims to help us determine whether AGENT-bot has intentions. First, intentions must always come before the action they motivate. Second, there is a strong link between intentional actions and reasons for acting. Theorists often describe this connection as one of identity, where intentions represent a reason for acting through a logical structure. This structure is typically a syllogism: the major proposition represents the agent's goal, and the minor proposition represents the agent's judgment on how to achieve it (Anscombe, 1957), (Davidson, 1963). Some theorists, such as Davidson, suggest that desires can serve as the major proposition and beliefs as the minor one, together forming an intention or intentional (Davidson, 1963), (Davidson, 1969), (Goldman, 1970), (Audi, 1986). Although this view is still widely accepted, some later philosophers argue against reducing intentions purely to desires and beliefs (Enç, 2003), (Bratman, 1987), (Bishop, 1989), (Mele, 1992). Nonetheless, it is generally agreed that intentions involve a ratiocination that justifies the action. In line with the SCA's first claim, this reasoning, which defines or constitutes intentions, must precede the action.

AGENT-bot does appear to conform to the SCA. Take, for example, an instance where AGENT-bot approaches and contacts a Triangle. This would qualify as an intentional action only if it is preceded by an intention that embodies a reason for making contact with the Triangle. Following the SCA framework, this reason could be represented by a syllogism; a major proposition indicating AGENT-bot's goal and a minor proposition representing its judgment on how to achieve it. In AGENT-bot's case, the major proposition could be a motivation to maximize its cumulative reward signals, while the minor proposition could be a learned judgment that contacting Triangles will yield rewards.

It's important to note that the SCA does not specify a precise definition of what constitutes an intention; instead, it identifies a type of action. The action must be preceded and informed by an intentional state or condition closely related to a reason for acting. In the example above, AGENT-bot's decision to contact the Triangle is preceded by reasoning it owns that support and explains the action. This shows that AGENT-bot aligns with the SCA.

6.2.2. Standard Theory of Agency.

The philosophy of action also defines a Standard Theory of Action. The standard theory of action is derived from but should not be conflated with the standard conception of action. The STA describes in causal terms what constitutes intentional action and therefore reasons that support intentional actions. The theory states that an intentional action is an action that is caused by the agent's right mental state/s, acting on or in response to the right event/s, in the right way (Schlosser, 2019). The right way in this case is an action that is rationalized from the agent's point of view as an appropriate means to approach occurrent events. In other words, the outcome of the action represents an appropriate goal for the agent given the occurrent events. Often the right mental states are defined in terms of beliefs and desires (Goldman, 1970), (Davidson, 1980), (Dretske, 1988). Therefore, in the standard theory of action, the reason for actions is quite frequently defined in terms of rational ends that are derived from the agent's mental states such as beliefs and desires interacting with or in response to occurrent events for the agent's end.

For example, consider an agent named Jimmy. Jimmy is thirsty (state), he desires to quench his thirst (desire), and he believes the bottle of water in the fridge will quench his thirst (belief). Jimmy drinks the water from the fridge (action). This action is considered intentional because it can be rationalized as a means to achieve Jimmy's goal, caused by his mental states interacting with current events (his thirst and belief about the water).

It must be noted that the standard theory of action does not commit to a desire-belief borne explanation of intentions. Some thinkers argue that intentions can be construed as irreducible mental states that inform intentional actions (Bratman, 1987), (Enç, 2003), (Bishop, 1989), (Mele, 1992). However, even in the case where intentions are construed as irreducible mental states, the standard theory of action holds. An intentional action is caused by the right mental states, in this case an irreducible intentional state, acting in response to or on the right events, in the right way.

From the standard theory of action, one can derive a standard theory of intentional agency. An entity has intentional agency if it has the right functional organization through which the right mental state/s can interact with the right events to cause the right action (Goldman, 1970) (Dretske, 1988), (Bratman, 1987), (Mele, 1992), (Davidson, 1980) (Brand, 1984), (Enç, 2003). The right action in this case being an action that can be rationalized as an end from the agent's point of view given occurrent mental states and events.

Does AGENT-bot conform to the STA?

The question as to whether AGENT-bot conforms to the STA is going to turn on whether one can identify the proper functional organization that would cause the appropriate events in the appropriate manner towards AGENT-bot's ends. The STA holds that an entity has the capacity to act intentionally if it has the proper functional organization to cause an appropriate event in an appropriate manner (Schlosser, 2019). This functional organization is instantiated by the right mental states and events (often stated in terms of desires, beliefs, and intentions) that cause the appropriate events in the appropriate manner. Agency, therefore, entails the establishment of the proper causal relationships between agent's states and events (Davidson, 1963), (Enç, 2003), (Bishop, 1989), (Dretske, 1988).

AGENT-bot does appear to have a proper causal relationship between its involved states and experiences and the actions it puts out. AGENT-bot seems to exhibit a clear causal relationship between its internal states, experiences, and resulting actions. Consider, for example, a scenario in which AGENT-bot approaches and contacts a Square. What states and occurrences might have led to this action?

1. AGENT-bot is driven by a motivation to maximize reward signals from its environment.
2. It experiences a qualitative state *(ad litteram)* indicating low energy, which triggers a negative reward signal.
3. Through associative learning, AGENT-bot has learned that contacting Squares reduces the negative reward signal.

4. AGENT-bot perceives an external thermic signal, represented as light in its internal sensory model, which identifies a nearby Square.
5. AGENT-bot then moves toward and makes contact with the Square.

This sequence illustrates how AGENT-bot's internal motivations, learned associations, and sensory perceptions work together to produce goal-directed actions.

In this case AGENT-bot does appear to have the right functional organisation to cause the appropriate event (movement towards and contact with a Square) in an appropriate way. At the risk of accusations of anthropomorphising, one might make this functional organization more visible by drawing reference lines between AGENT-bot's internal states and the language used to describe operative states in most action theories, i.e., beliefs, desires, and intentions.

1. AGENT-bot's **intention** is to maximize cumulative reward signals.
2. AGENT-bot is in a state of discomfort (negative reward signal paired to low-energy state) and **desires** to alleviate that discomfort.
3. AGENT-bot has a learnt **belief** that contact with a Square (or to put it in AGENT-bot's perspective - an item with a medium heat signature) will alleviate the discomfort.
4. AGENT-bot perceives a Square and forms an instrumental **intention** to contact a Square.
5. AGENT-bot takes the **action** to make contact with a Square.

The claim here is *not* that AGENT-bot has beliefs and desires identical to those of a human. The previous paragraph cautiously used terms from human action theory, such as desires, beliefs, and intentions, to describe AGENT-bot's states. This was done to underscore the functional similarity between AGENT-bot's internal states and those commonly used to explain human action processing, while fully acknowledging the differences. It might very well be the case that human specific action-processing states are different in how they present from other cognitive creatures too. What they have in common is functional similarity in accounting for intentional behaviour.

The point that AGENT-bot has its own action processing states that are differentiable, but functionally like humans is a critical point to note, because it might be tempting to think that AGENT-bot does not really perform its own ratiocination. The suspicion is that its involved action processing states only have meaning from a human perspective, and that AGENT-bot is not performing any reasoning of its own; rather it is the case that humans are simply projecting their own meaning onto AGENT-bot's involved states.

This is not the case. The rationalisation that informs AGENT-bot's actions are entirely its own and can be differentiable from its human associate's rationalizing even while they converge on the same goal. AGENT-bot's reasons for acting, just like the motivation for its human associate, is fundamentally geared towards lowering the temperature in the environment. While lowering the temperature is a concept that relates to objective features of the environment that are comprehendible by both AGENT-bot and its human associate, it means different things to both. For the human it means prolonging the demise of the environment and instrumentally leading to satisfying some motivating intrinsic pleasure, perhaps knowledge gained from studying the environment. For AGENT-bot lowering the temperature entails satisfying its own intrinsic reward signal. Both the human and AGENT-bot have reasons grounded in the same objective world, but their ratiocination is subjective to their unique perspective. They are labouring under their own unique reasons, even while these reasons converge on the same objective.

To further illustrate that AGENT-bot has its own reasoning processes that are distinct from those of its human handlers, consider the concept of an *Umgebung*, which refers to an umwelt as perceived by an outside observer (Cobley, 2010, p. 348). When a human operator observes AGENT-bot approaching and consuming a Square, they might rationalize this action by positing a major premise: AGENT-bot is hungry or low on battery (depending on whether they view AGENT-bot as a device or an agent). The minor premise would then be that Squares provide sustenance or a recharge.

However, these interpretations do not align with the concepts AGENT-bot uses in its reasoning. AGENT-bot's *Umwelt* does not recognize the shape of squares; instead, it perceives light signals that indicate minimal or safe heat signatures. Similarly, AGENT-bot does not experience hunger but instead detects a qualitative pattern associated with a negative reward signal. Therefore, AGENT-bot is not seeking out a Square to satisfy hunger or recharge; rather, it is pursuing medium heat signatures that will yield a reward to alleviate its negative reward signal.

This distinction is crucial because recognizing intentional actions involves understanding that an intentional action should be rationalized from the agent's own perspective, based on its own meaningful internal states. In the Standard Theory of Action (STA), the agent's perspective and experience are vital for grounding the rationalization process that explains its actions. The fact that AGENT-bot interprets its internal states and experiences and uses this information to determine appropriate actions demonstrates its conformity with the STA.

The Standard Theory of Agency describes an **Event-Causal framework for agency**. The Event-Causal framework covers accounts of agency that define agential actions in terms of event-causal relationships between the right agent states and events (Schlosser, 2019).

It would be difficult to deny AGENT-bot agency under the Event-Causal framework. One could certainly interpret AGENT-bot's actions by reducing them down to the causal effect of agent involving states and events. While the Event-Causal framework is the most dominant and widely accepted view on agency (Schlosser, 2019), there are other metaphysical views on agency, and it is worth exploring what implication they will have on AGENT-bot's agency.

6.2.3. AGENT-bot and other metaphysical frameworks for Agency.

The Agent-Causal Framework

The Agent-Causal framework objects to the reduction of actions in intentional agency to agent involving states and events. It asserts that the agent has an irreducible role in intentional agency that consists of exercising agent-causal power (Schlosser, 2019), (Lowe, 2008), (O'Connor, 2000), (Clarke, 2003). The main difference between Agent-Causal framework and Event-Causal frameworks is that the latter's causal chain singles out the agent involving states without regard for the agent. The Agent-Causal frameworks insist that the agent must be located in the causal chain of action as a persisting substance.

In identifying agential actions as actions with the right agent-causal history rather than the right event-causal history, the Agent-Causal frameworks object to the Standard Theory of Action (STA). However, they are in conformity with the Standard Conception of Action (SCA). The SCA states that agential action is preceded by a causal history that consist of intentions that are defined or connected to a reason for acting. The SCA does not commit to a particular account of intentions and therefore does not object to singling out a role for the agent as a persisting substance in the causal history of an action.

In being in conformity with the SCA, adherents of the Agent-Causal framework should have no interest in denying AGENT-bot agency. If the Agent-Causal theorists agree that AGENT-bot conforms to the SCA, what they will find objectionable is the Event-Causal interpretation of its actions. The task at hand would be to proffer an account of AGENT-bot's agency that traces the role of an agent as a persisting substance in the causal history of its agential actions. This might take the form of identifying a state that is definitive of the agent and therefore represents the agent functionally in the causal chain of action. In the case of AGENT-bot, this might be the inclination to maximize cumulative reward signals. This state is persistent and efficacious in all AGENT-bots action processing. Under this interpretation, the inclination to maximize reward signals is inseparable from AGENT-bot and therefore representative of AGENT-bot in the causal history of agency. Such an interpretation of AGENT-bots agency would satisfy an Agent-Causal framework.

The Volitionist Framework

A third metaphysical framework for intentional agency is called the volitionist approach. The volitionist approach rejects both the SCA and the STA. It refutes that agential actions should be defined by their causal history (Ginet, 1990), (McCann, 1998). It holds that all actions are fundamentally caused by volitions or acts of the will. These volitions are uncaused. Under this framework, you are an agent or possess agency if you are a thing or entity that has volitions.

It is not clear what the volitionist will make of AGENT-bot's agency. The problem here is that they have taken the very thing debates about intentional agency aspire to explain and rendered it beyond examination. Does AGENT-bot have intentional agency? Well, only if it has volitions. What are volitions? They are uncaused features of mental processes. The fact that the volitionist approach does not explain how volitions account for control of actions beyond simply hosting a volition, is a reason why volitionist approaches are widely rejected (Schlosser, 2019), (O'Connor, 2000, pp. 25-6), (Clarke, 2003, pp. 17-24).

One approach the volitionist might be tempted to take regarding AGENT-bot's agency is to state that volitions are unique to brain processes. However, if they took that stance against AGENT-bot, they would open themselves up to a much bigger challenge. They would have to explain why the brain, a physical entity, can host events that are exempt from the principle of causality.

6.2.4. Realist Theories of Agency.

The previous section briefly highlighted a realist position on intentional agency in the volitionist framework and concluded that it was too opaque to examine. However, one could take a realist stance about the STA too. This stance would question whether artificial systems can host representational mental states like those in the mind that appear to be essential in the causal history of agency under the STA. The argument here could be that intentional actions should only be ascribed to those representational states that arise from brain activity. The underlying assertion being that intentional agency is material specific. The representational states on which it relies are unique to the specific biochemistry of the mind.

Let it be assumed for arguments sake that one accepts the position that the designation of intentional actions should be ring fenced for actions caused by mental states in the mind. It should be kept in mind that defining intentional agency is to explicate a functional process. It describes the capacity to process intentional actions. What is interesting about intentional agency is how it functions, or how this capacity is instantiated. One might accept on the grounds of the argument that intentional agency is specific to brain material and that AGENT-bot on those terms has no intentional agency. However, this would not necessarily mean that

AGENT-bot does not function like an intentional agent. AGENT-bot has representational states with meaning that are efficacious in the causal history of its actions. These representational states are functionally analogous to the mental states in the mind that account for intentional actions. The realist will have to proffer an explanation for what the specific biochemistry of the mind contributes to the functional process of agency and tell us where these processes are missing in AGENT-bot. Short of this, they would have to agree that AGENT-bot is functionally agential. If there is agreement that AGENT-bot functions like an intentional agent, or is intentional agency adjacent, the realist is still well within their rights to continue to reserve the word 'intentional actions' for mind stuff. However, they would have to agree that their objection is a quibble about semantics.

The far more parsimonious stance to take on whether artificial systems can have representational states like those in the mind is an instrumentalist position. The instrumentalist position is best encapsulated in Daniel Dennet's intentional stance. Dennet suggests that we should attribute agency to a system to the extent that taking an intentional stance towards the system is useful in explaining and predicting its behaviour (Dennet, 1987, p. 17), (Shoemaker, 1990). According to the intentional stance we should not shy away from defining AGENT-bots representational states in the language of intentional mental states (e.g., beliefs, desires, intentions), if attributing these states to AGENT-bot can explain or predict its rational behaviour.

The philosophical accounts of intentional agency favour that AGENT-bot's is capable of intentional action. There appears to be not enough coherent ground on which to deny that AGENT-bot is performing intentional actions.

6.3. Conclusion: Hephaestus Golden maidens and agency.

Hephaestus' golden maidens are described as complex contraptions. In comparison, AGENT-bot is a quite simple contraption. It's important to remember that AGENT-bot represents a minimal viable product of artificial general intelligence (AGI). However, the example of AGENT-bot illustrates that if Hephaestus had employed some version of the Analog Principle (AP) to allow his golden maidens to derive meaning from the world and some version of the Internal Model Principle (IMP) to enable them to take ownership of their actions, it would be hard to argue against their capacity for taking intentional agency that is functionally comparable to human autonomy.

Chapter 7: Conclusion

> *"Science is progressing at such a rapid rate that when you make a prediction and think you are ahead of your time by 100 years you may be ahead of your time by 10 at most."*
>
> Leo Szilard (LIFE Magazine, 1961, p. 79).

7.0. Ground covered in this thesis.

Artificial General Intelligence is a poorly defined concept. This poverty in specificity has added to the difficulties in realizing an AGI artefact. It is not unfathomable to discover that different leading experts at the cutting edge of the research field might be labouring under different concepts of artificial general intelligence.

As an example of the diversity in concepts that have been attached to AGI, consider that there seems to be broad agreement that the muse for the AGI project is human intelligence. The anticipated AGI will have intelligence that is comparable to human intelligence in some important way. Yet this anticipated comparability to human intelligence varies depending on which expert you ask. Some have claimed that the comparability should contrast narrow intelligence by matching human intelligence's generality (Goertzel & Pennachin, 2007), others have claimed that AGI would have the capacity to perform any intellectual task that a human can do (Goertzel, 2014), others that the comparability should be tagged to human's ability to apply learning from one area to another (Shanahan, 2015).

To compound the matter, disagreements like these are extended to the definitions of the key words in artificial general intelligence. The terms 'intelligence' and 'general' attract similarly varying and competing interpretations in the research field. Because the research field is poorly defined, it is accommodative to interpretations and diverse approaches. The result is that different AGI R&D projects might be approaching the task at hand under different methods and working towards different goals.

Therefore, to tackle the subject of AGI in this thesis, choices had to be made. The definition of AGI adopted and defended in this thesis was labelled the Grand Ambition. It was so called because it conformed to the stated overarching ambitions of the architects of the AI research field, who when describing the field of research they wished to pursue, defined it as a quest to realize machines that had intelligence that is functionally comparable to human intelligence (McCarthy, et al., 1955).

Chapter two interpreted the Grand Ambition's mandate as the ambition to realize an AGI artefact that has problem solving competencies that are general in the sense that the artefact can self-acquire solutions to familiar, differently contextualized and new problems in its native environment. These solutions are output in the form of actions or behaviours. The artefacts problem solving competencies should be comparable to human intelligence in a way that is not strictly defined by its capacity or capability, but rather its functional processes.

With whichever definition of artificial general intelligence that one adopts, there are a few anticipated properties of artificial general intelligence that are unlikely to draw much disagreement. It is broadly accepted that a realized AGI will possess the ability to understand, learn, and apply knowledge across a wide range of tasks and it would do so by acting on its own.

Even if one were to limit their research interest in AGI to these 3 key features, broad understanding, learning and autonomy, that would still constitute a very wide field of study. This thesis has limited itself to a discussion of only one of these features, the principles that realize the desired autonomy in AGI.

The argument in this thesis has been that if the ambition of AGI is to realize functional comparability to the human mind, then a realized AGI must have the same autonomy as the mind. The thesis has proceeded under a computational theory of the mind. It holds that the mind is indeed doing something akin to computation. One thing that sets the mind apart from artefactual computing is that the mind manages its processes without outside agency and in a manner that is wholly creditable to itself. AGI in its quest for functional comparability with the mind, should seek to replicate the way the mind manages its computational self-sufficiency.

The thesis has proposed and defended two principles that can support computational self-sufficiency that is functionally comparable to the self-sufficiency of the mind. The Analog Principle describes how a computational system can self-ground underived meaning into its computational items, and the Internal Model Principle describes how a computational system can realize internal states or processes that are creditable with self-control. The thesis showed that together the two principles can be used to ameliorate computational systems with the self-sufficiency to realize the type of intentional actions that is like the autonomy supported by the mind.

7.1. Anticipating future research.

In being inspired by the mind, artificial general intelligence research can benefit from the expansive work done in the philosophy of mind. It would have been unwieldy to attempt a complete discussion of the relevant philosophy of mind topics that would have been informed, implicated or impacted by the meandering path the thesis cut through the canon.

Even with our interest only limited to one anticipated characteristic of AGI, the topic of autonomy, certain sacrifices had to be made in the interest of respecting the word limit of this piece of work. For example, the discussion around the implications of qualitative experiencing giving rise to a body schema in chapter 5 hinted at theories in 4E cognition. There was a risk that exploring this lead would have taken us too far astray from the direction our narrative was taking.

The thesis had to make several of these sacrifices. It cut a horizontal path through the vertical topics in the philosophy of mind canon. For example, the thesis adopted a computational theory of mind as more or less a given, it assumed a functional view of intelligence, the analog principle described a causal theory of mental representation, the internal model principle proposed a mental model theory, and the self-directedness achieved from the AP and IMP was interpreted under an instrumentalist position that is best encapsulated in Daniel Dennet's intentional stance (Dennet, 1987, p. 17). There is a potential thesis to be articulated at every point these decisions were taken. The thesis had to leave that work for others because there was no space in this work, and it did not fit in the narrative.

The thesis also limited itself to addressing the problem of AGI autonomy. There are other consensus properties of AGI that the thesis did not address. AGI is anticipated to have broad understanding and learning. The thesis has been silent on the challenges that abound in engineering for these properties. However, these problems have been central to the research field in both industry and academia. Billions of dollars have been invested into AGI research. Some frontier technologies are showing promise. For example, liquid neural nets appear to evidence progress towards algorithms that are not domain restricted (Hasani, et al., 2020). Neuromorphic computing appears to evidence improvements towards brain-like computational efficiency (Rathi, et al., 2023). LLMs are registering great improvements in matching human performance in benchmark tests (Aschenbrenner, 2024). AGI R&D appears to be advancing on several fronts.

However, despite the progress and broad optimism, the finishing line is not obvious, and neither is it given. There appears plenty of work to be done before we can realize an artefact that thinks and solves general problems in a native environment in a way that is functionally comparable to a human being. There is no guarantee that the progress we have realized to date is a guarantor for success.

With expectations tempered, it is worth reviewing what lays on the other side of the optimism spectrum.

7.2. Leo Szilard’s leap forward.

In the 1930s, a book by the remarkably prescient science fiction author, H.G. Wells, called The World Set Free (Wells, 1914) ignited the popular imagination on a topic that had until that point been the reserve of theoretical discussions in ivory towers. H.G Wells had authored the book in which he discussed the harnessed release of energy from atoms almost 20 years earlier. He had predicted that such technology could be leveraged in warfare. Perhaps it was the imminence of a continent-wide European war that loomed over much of the early 1930’s that had brought his book and ideas to fore of the public’s imagination two decades after it had been written.

H.G. Wells’ writing had been influenced by science. He had kept abreast with the work of atomic physicists of his day such as William Ramsay, Ernest Rutherford, and Frederick Soddy. The science had long discussed the slow but steady release of energy in radioactive decay.

In the 20 years that Wells’ book had attracted little notice from the public, the science had progressed. The mainstream thinking on the topic was that while the total amount of energy released from radioactive decay was substantial, it was released at such a ponderous rate that the harvestable amounts were negligible. In a lecture in 1933, Lord Ernest Rutherford, the preeminent expert on the topic, addressed the public’s fascination with harnessing the energy locked in atoms. With the backing of the best knowledge of the day, he dismissed the idea as fanciful, going as far as to call it ‘moonshine’ (Rhodes, 1986, p. 27).

On September the 12th of the same year, Leo Szilard, a Hungarian immigrant, was mulling over the article in the Times that reported on Lord Rutherford’s lecture. As he stood waiting on a traffic light that still stands today at the point where Southampton Row meets Russell Square, he conceived of the idea of a nuclear chain reaction. He realized that if one found the right elements, their neutrons could be prompted to produce a nuclear reaction that produced more neutrons, that in turn would continue splitting into more neutrons by the same reaction. This chain reaction would be self-sustaining and would release tremendous amount of energy with little input (Lanouette & Silard, 1992).

Just as Lord Rutherford, the man widely recognized as the father of nuclear physics, was dismissing the idea that harnessing energy from splitting an atom as fantasy, the problem was solved in a remarkable moment of insight on a grey morning on the streets of Bloomsbury.

7.3. Lessons for AGI research.

Very similarly to the state of atomic research at the point of Leo Szilard’s conceptual breakthrough, AI research is seeking a great conceptual leap forward to unlock AGI. It is likely the case that the gap to be covered towards realizing AGI could be one insight away. If this is the case, this insight is most likely waiting to be discovered away from mainstream AGI research.

This is not because mainstream AGI research is particularly flawed, it is rather that the nature of most types of mainstream scientific research develops in a bubble that builds incrementally from its own existing knowledge base. For the most part, this process works well enough to advance science. However, it can be a poisoned chalice if misconceptions enter the existing knowledge base, or even worse, in cases where there are significant blind spots in fundamental knowledge. Errors in the existing knowledge base might eventually be shown to be wrong and corrected, but one cannot correct what is not known.

Lord Rutherford was not speaking out of turn when he expressed his pessimism about the prospects of harnessing the energy from splitting the atom. His position was supported by the latest research. Some of the most cutting-edge research had been done by his students, John Cockcroft and Ernest Walton who in 1932 had managed to split lithium atoms into alpha particles (Lanouette & Silard, 1992, pp. 131-2). It was on the basis of such cutting edge developments in the field that Lord Rutherford concluded that while splitting the atom was academically interesting, it was not a practical source of energy. Lord Rutherford, because of his privileged position, had access to an extensive canon of work on atomic physics to draw from in drawing conclusions on matters in the subject. However, that privileged position, would have done him little good in thinking outside of the box to discover brand new insights. In fact, it is arguable that it would have done him a disservice.

Leo Szilard, on the other hand, was able to attack the problem unencumbered by mainstream anchoring bias, and with less risk of the reputational harm associated with failure.

This not to say that Leo Szilard was not a genius, because he was by every measure. In addition to his contributions to physics, Szilard was an engineer who with Albert Einstein helped co-develop the Einstein-Szilard refrigerator. He was also a biologist of great renown, a field in which his contributions include the discovery of feedback inhibition in enzymes and in which he made contributions that made the cloning of human cells possible. In addition to his outsider perspective in atomic physics research, this multi-subject competency most likely gave Szilard an advantage in being able to make his breakthrough towards splitting the atom. By applying a systemic view enriched with knowledge from neighbouring disciplines, Szilard was more likely to be able to apply solutions and insights from different domains to inform atomic physics.

There are lessons to be learned from Leo Szilard's conceptual leap forward that could inform AGI research. One important lesson is that the research field must guard against an anchoring bias. An anchoring bias is a tunnel vision that favours information that is already known. The AGI research field is especially susceptible to an anchoring bias because of the success that has been achieved in narrow AI research. It is tempting to think that an AGI will emerge in some form or flavour by continuing down the same path that has achieved success in ANI research.

The opinion that AGI would emerge from incremental improvements in ANI had started to wane, but it has since re-emerged with conviction recently as LLM foundational models have been shown to be effective when applied to a wide range of tasks with just a little fine tuning and no adjustment to their underlying architecture. Some in the research field now believe that these large foundational models, popularly known as generative AI, are the first flickers of artificial general intelligence (Bubeck, et al., 2023). Billions of dollars are currently being invested in expanding the performance of large foundational models partly because the case has been made that they are a legitimate path to AGI. If indeed this is the case, then the optimism and investment is warranted. However there remains a significant risk that this might be an extremely expensive case of anchoring bias.

AGI research has experienced previous instances of misplaced optimism. Throughout its history, researchers in the field have consistently held high hopes. To illustrate, in the 1950s, Alan Turing predicted the development of AGI by the year 2000. A decade later, Arthur C. Clarke extended this prediction to 2001 (Deutsch, 2012). Now, more than two decades beyond those anticipated due dates, AGI has not materialized. Historical patterns reveal that periods of optimism in the field are often followed by a decline in interest and funding for AI research, leading to what is commonly referred to as AI winters. It's clear that optimism alone is not a reliable indicator of progress in the research field.

The problem is likely the case that the gap yet to be covered towards AGI relates to a matter of foundational importance in understanding the nature of general intelligence. If this is indeed the case, then incremental improvements that build on existing knowledge within AI might do well enough to create the impression of progress, but it will not realize an AGI. What is needed is a fresh perspective, a view unencumbered by entrenched approaches and ideas. If a fundamental gap within the research field remains unaddressed, mere incremental advancements on existing knowledge, no matter how inspired, will not bridge the gap. The solution lies in a paradigm shift that either enhances, broadens, or entirely redefines the foundational understanding in the field.

This highlights a second lesson that can be gleaned from Szilard's breakthrough. Szilard was a polymath. He was able to draw from neighbouring disciplines and their concepts to broaden his theorising in atomic physics. If indeed the roadblock to AGI is a poverty in the understanding of the concept of general intelligence, then the research field will benefit from exploring ideas and advancements in neighbouring disciplines that touch upon related subjects.

An insightful area to explore is the philosophy of mind, which delves into the ontology and nature of the mind. This field encompasses a rich body of work that addresses diverse aspects of the mind, including knowledge, consciousness, cognition, phenomenal experiences, action theory, emotional states, perceptions, rationality, free will, reasons as causes, etc. These topics seem highly pertinent to constructing a comprehensive understanding of general intelligence. If the problem that is handicapping AGI research is an incomplete understanding of the mind, then David Deutsch is right in arguing that philosophy will be the key that unlocks artificial intelligence (Deutsch, 2012).

A thoughtful cross-functional approach will yield mutual benefits. Mainstream philosophy of mind stands to be enriched by progress in philosophy of AI Theory. Philosophical concepts proven effective in AI applications will significantly enhance theorizing and understanding within the philosophy of mind. In this context, AI theory could play a role for philosophy of mind similar to the role experimental physics plays in complementing theoretical physics. It will not only strengthen ongoing research within philosophy of mind with empirical evidence, but it will generate fresh insights and open new avenues for exploration.

7.4. A final lesson for the Philosophy of AI Theory.

There is one more crucial lesson to be gleaned from Leo Szilard’s breakthrough. Szilard had read H.G. Wells, and he was aware from the warnings in Wells’ The World Set Free, that the technology that harvested energy from atoms could potentially be put to nefarious use. Influenced by Wells' cautionary tale, when Szilard applied for a patent for his idea, he deliberately assigned it to the Admiralty, the department overseeing the Royal Navy. He thought that the precautionary step would prevent the technology from entering the public domain where it could be exposed and used by nefarious actors. However, Wells’ prophesy was not to be denied. In just over a decade, Szilard’s idea, complemented and built on with the work of various other geniuses such us Edward Teller, J. Robert Oppenheimer, Neils Bohr, Lise Meitner, Enrico Fermi, John Von Neumann, and others, resulted in the detonation of the first nuclear weapon on July 16th, 1945. A month later this weapon was used in war.

The use of nuclear energy in a bomb highlighted the existential threat posed by the technology. It prompted a worldwide demand for retrospective regulation. In response, eighty-one countries came together in 1957 to establish the International Atomic Energy Agency (IAEA). This global cooperative governance body is tasked with regulating and establishing rules and policies to ensure the safe utilization of atomic energy. While it has its flaws, the IAEA has to date managed to effectively mitigate the existential risks associated with nuclear technology while ensuring its ongoing peaceful applications.

Like nuclear technology, the potential risks associated with advanced AGI have consistently been emphasized in its popular representation. Many concerns raised in the portrayal of advanced AGI have been validated by convincing arguments in academia. There is substantial evidence indicating that the irresponsible development of AGI could present certain grave risks to humanity. Given that some of these risks might be existential, the regulation and governance of AGI must emulate the global collaborative efforts that has so far, successfully mitigated the threat of a nuclear apocalypse. We should aspire to establish a global and collaborative governance framework to ensure the responsible development and to reserve the use of AGI for the benefit of all humanity.

Fortunately, there is evidence to suggest that lessons have been learned from the tortured journey to regulating atomic energy. There are concerted efforts globally to establish regulations, policies, and governance structures for AI. Many of these efforts have made aspirations to global collaboration although tangible collaborative action remains thin on the ground.

Nevertheless, the most compelling evidence that we have learned valuable lessons since Leo Szilard's unsuccessful attempt at safeguarding his breakthrough from nefarious use, is the continuing maturity and integration of AI ethics in AI theory. Ethical considerations based on research have made substantial inroads into mainstream AI research. Work in ethics influences policies, frameworks, practices, and regulations in AI R&D and industry. While progress in the ethics of AI has been notable, there remains much work to be done. It is imperative not only to firmly establish ethics as an integral component of mainstream AI theory but also to ensure that the ethics of AI as a research discipline advances alongside or ahead of the development of AI itself. This is the most important contribution that philosophy of AI theory can make to the Grand Ambition of AI theory.

References.

Agamben, G., 2003. The Open: Man and Animal. In: K. Attel, ed. *The Open: Man and Animal.* Stanford California: Stanford University Press.

Albus, j. S., 1991. Outline for a theory of intelligence. *IEEE Trans. Systems, Man,* p. 473-509.

Aleksander, I. & Hanna, K. F., 1976. *Automata Theory: An Engineering Approach.* s.l.:Edward Arnold.

Amsel, A., 1992. Confessions of a neobehaviorist. *Integrative Physiological and Behavioral Science,* 4(27), pp. 336-346.

Anscombe, E. G., 1957. *Intention.* Oxford: Basil Blackwell.

Aristotle, 2000. Nicomachean Ethics. In: R. Crisp, ed. *Cambridge Texts in the History of Philosophy.* Cambridge: Cambridge University Press, pp. 37-59.

Armstrong, D., 1981. *What is consciousness?.* Ithaca, New York: Cornell University Press.

Aschenbrenner, L., 2024. *Situational Awareness: The Decade Ahead.* [Online]
Available at: https://situational-awareness.ai/
[Accessed 04 08 2024].

Ataria, Y., Tanaka, S. & Gallagher, S., 2021. *Body Schema and Body Image: New Directions.* Online ed. Oxford: Oxford Academic.

Audi, R., 1986. Acting for a Reason. *Philosophical Review ,* 95(4), pp. 511-546.

Bach-Y-Rita, P. et al., 1969. Vision substitution by tactile the image projection. *Nature. 221 (5184): 963-964.,* p. 963-964..

Baum, F. L., 1914. *Tik-Tok of Oz.* Chicago: Reilly & Lee.

Baum, L. F., 1907. *Ozma of Oz.* Chicago: Reilly & Britton.

Baum, S., Goertzel, B. & Goertzel, T. G., 2011. How Long Until Human-Level AI? Results from an Expert Assessment. *Technological Forecasting & Social Change, 2011, 78(1),* pp. 185-195.

Bayne, T. & Pacherie, E., 2007. Narrators and comparators: the architecture of agentive self-awareness. *Synthese ,* Volume 159, pp. 475-491.

Bishop, J., 1989. *Natural Agency: An Essay on the Causal Theory of Action.* Cambridge: Cambridge University Press.

Blakemore, S. J. & Frith, C., 2003. Self-awareness and action. *Current Opinion in Neurobiology,* 13(2), pp. 219-224.

Blanke, O. & Metzinger, T., 2009. Full-body illusions and minimal phenomenal selfhood. *Trends in Dognitive Science,* 13(1), pp. 7-13.

Block, N., 1978. *Troubles with Functionalism.* Minneapolis.: University of Minnesota Press.

Blum, C., Winfield, A. F. T. & Hafner, V. V., 2018. Simulation-Based Internal Models for Safer Robots. *Frontiers in Robotics and AI,* Volume 4.

Bongard, J., Zykov, V. & Lipson, H., 2006. Resilient machines through continuous self-modeling. *Science,* 314(5802), pp. 1118-21.

Botvinick, M. & Cohen, J., 1998. Rubber hands 'feel' touch that eyes see. *Nature 391 no.756,* pp. 10-1038.

Brand, M., 1984. *Intending and Acting: Toward a Naturalized Action Theory.* Cambridge, MA: MIT Press..

Bratman, M. E., 1987. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press..

Braun, N. et al., 2018. The Senses of Agency and Ownership: A Review. *Frontiers in Psychology ,* Volume 9.

Brenner, E. D. et al., 2006. Plant neurobiology: An integrated view of plant signaling. *Trends in Plant Science,* 11(8), pp. 413-419.

Bubeck, S. et al., 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712,* Volume 5.

Camacho, E. F. & Bourdons, A. C., 2004. *Model Predictive Control.* New York: Springer.

Cartwright, D. E., 2000. *Tides. A Scientific History.* Cambridge.,UK: Cambridge University Press.

Chalmers, D., 2010. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies,* 17(9-10), p. 7-65.

Chekhchoukh, A., Vuillerme, N., Payan, Y. & Glade, N., 2013. Effect of saccades in tongue electrotactile stimulation for vision substitution applications. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.,* Volume Annual International Conference, pp. 3543-3546.

Chisholm, R., 1966. Freedom and Action. In: *Freedom and Determinism.* New York: Random House, p. 11-44.

Christian, K. M., 2010. Cerebellum: Associative Learning. *Encyclopedia of Behavioral Neuroscience,* pp. 242-248.

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

Clarke, R., 2003. *Libertarian Accounts of Free Will.* Oxford: Oxford University Press..

Clarke, R., 2010. Skilled Activity and the Causal Theory of Action. *Philosophy and Phenomenological Research,* 80(3), p. 523-550.

Cobley, P., 2010. *The Routledge Companion to Semiotics.* London and New York: Routledge. p. 348..

Collins , 2023. *Sentience.* [Online]
Available at:
https://www.collinsdictionary.com/us/dictionary/english/sentience#:~:text=Definition%20of%20'sentience'&text=1.,not%20involve%20thought%20or%20perception

Conant, R. C. & Ashby, R. W., 1970. Every Good Regulator of a System must be a Model of that System. *International Journal of System Science,* 1(2), pp. 89-97.

Craik, K., 1943. *The Nature of Explanation.* Cambridge: Cambridge University Press.

David, N., Newen, A. & Vogeley, K., 2008. The “sense of agency” and its underlying cognitive and neural mechanisms. *Consciousness and Cognition Volume ,* 17(2), pp. 523-534.

Davidson, D., 1963. Actions, Reasons, and Causes. *The Journal of Philosophy,* 60(23), pp. 685-700.

Davidson, D., 1969. How is Weakness of the Will Possible. In: *Moral Concepts.* Oxford: Oxford University Press .

Davidson, D., 1980. Agency. In: *Essays on Actions and Events.* Oxford: Clarendon Press., pp. 43-61.

de Vignemont, F., 2010. Body Schema and Body Image-Pros and Cons. *Neuropsychologia,* 48(3), p. 669–680..

de Vignemont, F., 2011. Embodiment, ownership and disownership. *Consciousness and Cognition,* 20(1), pp. 82-93.

de Vignemont, F., Pitron, V. & Alsmith, A. J. T., 2021. What is the body schema?. In: Y. Ataria, S. Tanaka & S. Gallagher, eds. *Body Schema and Body Image: New Directions.* Oxford: Oxford University Press.

Dennet, D., 1987. *The Intentional Stance.* Cambridge, MA: MIT Press..

Dennett, D., 1978. Why Not the Whole Iguana?. *Behavioral and Brain Sciences,* Volume 1, pp. 103- 104.

Dennett, D., 1990. Quining Qualia. In: *Mind and cognition : a reader.* Cambridge, Massachusetts, USA: Basil Blackwell, pp. 519-549.

Dennett, D., 1992. The Self as a Center of Narrative Gravity. In: *Self and Consciousness: Multiple Perspectives.* Erlbaum: Hillsdale, NJ, pp. 103-116.

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

Deutsch, D., 1985. Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences,* 400(1818), pp. 97-117.

Deutsch, D., 2012. Philosophy will be the key that unlocks artificial intelligence. *The Guardian*, 3 October.

Dickinson, A., 1985. Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society,* 308(1135), pp. 67 - 78.

Dretske, F., 1981. *Knowledge and the Flow of Information.* Cambridge, MA: MIT/Bradford Press: MIT/Bradford Press.

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

Dreyfus, H., 1978. *What Computers Cant Do: The Limits of Artificial Intelligence.* New York; New York: HarperCollins.

Dreyfus, H. L., 2012. A History of First Step Fallacies. *Minds and Machines ,* Volume 22, pp. 87 - 99.

Dung, L., 2024. Understanding Artificial Agency. *The Philosophical Quarterly,* Volume 7.

Edelman, S., 2002. Constraining the neural representation of the visual world. *Trends in Cognitive Science,* 6(3), pp. 125-131.

Enç, B., 2003. *How We Act: Causes, Reasons, and Intentions.* Oxford: Oxford Univeristy Press.

Epoch AI, 2024. *epochai.org.* [Online]
Available at: https://epochai.org/data/notable-ai-models
[Accessed 04 08 2024].

Fay, J. W., 1939. *American Psychology Before William James.* New Brunswick, N. J.: Rutgers University Press.

Field, H., 2001. *Truth and the Absence of Fact.* Oxford:: Clarendon Press.

Fitzgerald, M., Boddy, A. & Baum, S., 2020. *2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy.,* s.l.: Global Catastrophic Risk Institute Technical Report 20-1.

Fodor, J., 1975. *The Language of Thought.* New York: Thomas Y. Crowell.

Fodor, J., 1975. *The Language of Thought.* New York: Crowell.

Fodor, J., 1980. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Sciences,* 3(1), p. 63-109..

Fodor, J., 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge: : MIT Press.

Fodor, J., 1990. *A Theory of Content and Other Essays.* Cambridge, MA: : MIT Press..

Fogel, D. B., 1995. *Evolutionary computation: toward a new philosophy of machine intelligence.* 3 ed. Hoboken, New Jersey: IEEE Press.

Fogel, D. B., 1995. Review of computational intelligence: Imitating life.. *Proc. of the IEEE, 83(11),.*

Forch, V. & Hamker, F. H., 2021. Building and Understanding the Minimal Self. *Frontiers in Psychology,* 26(12).

Ford, M., 2018. *Architects of Intelligence.* Birmingham UK: Packt Publishing.

Francis, B. A. & Wonham, W. M., 1976. The Internal Model Principles of Control Theory. *Automatica,* 12(5), pp. 457-465.

Freeth, T. et al., 2006. Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. *Nature. ,* 444(7119), pp. 587-591.

Freeth, T., Jones, A., Steele, J. M. & Bitsakis, Y., 2008. Calendars with Olympiad display and eclipse prediction on the Antikythera Mechanism. *Nature ,* 454(7204), pp. 614-617.

Frege, G., 1956. The thought: A logical inquiry. *Mind,* 65(259), pp. 289-311.

Friberg, T. R. et al., 2011. "Seeing" With Your Tongue -- Sensory Substitution Using A Simple Alternative To The Retinal Chip. *Investigative Ophthamology and Visual Science,* 52(14).

Gallagher, S., 1986. Body image and body schema: A conceptual clarification. *The Journal of mind and behavior,* 7(4), p. 541-554.

Gallagher, S., 2000. Philosophical Conceptions of the Self: Implications for Cognitive Science. *Trends in Cognitive Sciences. ,* 4(1), pp. 14-21.

Gallagher, S., 2007. The Natural Philosophy of Agency. *Philosophy Compass,* 2(2), pp. 141-357.

Gallagher, S., 2020. Action, Intention, and the Sense of Agency. In: S. Gallagher, ed. *Action and Interaction .* s.l.:Oxford Academic, pp. 42-66.

Gallagher, S. & Zahavi, D., 2012. *The Phenomenological Mind..* New York: Routledge.

Gentner, D. D., 1989. The Mechanism of Analogical Learning. In: S. Vosniadou & A. Ortony, eds. *Similarity and Analogical Reasoning, Cambridge: Cambridge University Press..* Cambridge:: Cambridge University Press, pp. 199-241.

Gibson, J. J., 1966. *The Senses Considered as Perceptual Systems.* London: Allen and Unwin.

Gibson, J. J., 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin Harcourt (HMH).

Gillett, C., 2007. A Mechanist Manifesto for the Philosophy of Mind: The Third Way for Functionalists. *Journal of Philosophical Research, invited symposium on "Mechanisms in the Philosophy of Mind", vol.32,* pp. 21-42.

Ginet, C., 1990. *On Action, Cambridge.* Cambridge: Cambridge University Press..

Glock, H.-J., 2019. Agency, Intelligence and Reasons in Animals. *Philosophy,* 94 (4), pp. 645 - 671.

Goertzel, B., 2014. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence,* 5(1).

Goertzel, B. & Pennachin, C., 2007. *Artificial General Intelligence.* New York: Springer Berlin Heidelberg.

Goldman, A., 1970. *A Theory of Human Action.* Englewood Cliffs: NJ: Prentice-Hall..

Grace, K. et al., 2018. When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research ,* Volume 62, pp. 729-754.

Graziano, M., 2018. *The Spaces Between Us: A Story of Neuroscience, Evolution, and Human Nature..* New York: Oxford University Press.

Graziano, M. & Botvinick, M., 2002. How the brain represents the body: insights from neurophysiology and psychology. In: W. Prinz & B. Hommel, eds. *Common Mechanisms in Perception and Action: Attention and Performance Volume XIX.* New York: Oxford University Press, pp. 136-158.

Graziano, M. S., 2019. *A scientifc theory of subjective experience.* New York: W.W. Norton & Company.

Graziano, M. S. A., 2017. The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *Frontiers in Robotics and AI,* Volume 4.

Grice, R. G., 1957. Meaning. *The Philosophical Review,* Volume 66, p. 377-388..

Hafner, V. V., Loviken, P., Villalpando , P. A. & Schillaci , G., 2020. Prerequisites for an Artificial Self. *Frontiers in Neurorobotics ,* 14(5).

Haggard, P., 2017. Sense of agency in the human brain. *Nature Reviews Neuroscience,* Volume 18, pp. 196-207.

Hamilton, A. & McBrayer, J., 2020. Do Plants Feel Pain. *Disputatio,* 12(56), pp. 71-98.

Harnad, S., 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena,* 42(1-3), pp. 335-346.

Harper, D., 2021. *Sentient.* [Online]
Available at: https://www.etymonline.com/word/sentient

Hasani, R. et al., 2020. *Liquid Time-constant Networks,* Virtual (www.aaai.org): Association for the Advancement of Artificial Inteligence .

Haugeland, J., 1985. *Artificial Intelligence: The Very Idea..* Cambridge, MA: MIT Press.

Haugeland, J., 1985. *Artificial intelligence: The very idea.* Cambridge, MA: MIT Press, Bradford Books..

Haugeland, J., 2002. Syntax, semantics, physics. In: J. M. Bishop & J. Preston, eds. *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence..* London: Oxford University Press.

Haven, J., 1858. *Mental Philosophy: Including the Intellect, Sensibiliries and Will.* Boston: Gould and Lincoln.

Henry, M., 1973. Philosophie et phénoménologie du corps (1965). In: G. Etzkorn, ed. *The Essence of Manifestation.* The Hague: Martinus Nijhoff.

Hernández-Orallo, J., 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence.* Cambridge, United Kingdom: Cambridge University Press.

Hernández-Orallo, J. et al., 2021. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific Reports,* Volume 11, p. 22822.

Herstein, I. N., 1964. *Topics in Algebra.* s.l.:Ginn and Company.

Hilgard, E. R., 1980. The Trilogy of the Mind: Cognition, Affection and Conation. *Journal of the History of the Behavioral Sciences ,* 1(16), pp. 107-117.

Hoffman, M. J. et al., 2010. Body Schema in Robotics: A Review. *IEEE Transactions on Autonomous Mental Development,* Volume 2, pp. 304-324.

Holland, J. M., 2004. *Designing Autonomous Mobile Robots: Inside the mind of an intelligent machine.* 1st ed. s.l.:Newnes; Elsevier Inc..

Homer, 1924. *The Illiad; translated by A.T. Murray.* Cambridge MA: Harvard University Press.

Horgan , T. & Tienson, J., 2002. The Intentionality of Phenomenology and the Phenomenology of Intentionality. In: *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: Oxford University Press, 520-33., p. 520-33..

Husserl, E., 2019. Erste Philosophie II 1923-24. In: S. Luft & T. M. Naberhaus, eds. *First philosophy : lectures 1923/24 and related texts from the manuscripts (1920-1925).* Dordrecht. The Netherlands: Springer, p. 33.

Hutter, M., 2000. A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv:cs.AI/0004001.*

Hutter, M., 2007. Universal algorithmic intelligence: A mathematical top→ down approach. In: *Artificial general Intelligence. Cognitve Technologies.* Berlin, Heidelberg: Springer, pp. 227-290.

Iversen, P. A., 2018. The Calendar on the Antikythera Mechanism and the Corinthian Family of Calendars. *Hesperia.,* 86(1), p. 130.

James, W., 1890. *The Principles of Psychology.* New York: Henry Holt and Company.

Johnson-Laird, P., 2004. The History of Mental Models. In: *Psychology of Reasoning: Theoretical and Historical Perspectives.* New York: Psychology Press, pp. 179-213.

Johnson-Laird, P. N., 1986. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* United Kingdom: Harvard University Press.

Kant, I., 1781 [1998]. Critique of Pure Reason. In: P. Guyer & A. W. Wood, eds. *Critique of Pure Reason.* Cambridge: Cambridge University Press.

Kant, I., 2023. Critique of Judgement (1888). In: J. Wason, ed. *The Philosophy of Kant: As Contained in Extracts from His Own Writing.* New Edition ed. Glasgow: Maclehose, Jackson, pp. 307-323.

Kenton, Z. et al., 2023. Discovering Agents. *Artificial Intelligence,* Volume 322.

Kolarik, A. J., Scarfe, A. C., Moore, B. C. J. & Pardhan, S., 2017. Blindness enhances auditory obstacle circumvention: Assessing echolocation, sensory substitution, and visual-based navigation. *PLOS One,* 12(4).

Kolb, B. & Whishaw, I. Q., 2001. *An introduction to brain and behavior.* New York: Worth.

Kull, K., 2010. Umwelt. In: *The Routledge Companion to Semiotics..* London: Routledge., pp. 348-349.

Kurzweil, R., 2005. *The Singularity Is Near: When Humans Transcend Biology.* London: Viking.

Lanouette, W. & Silard, B., 1992. *Genius in the Shadows: A Biography of Leo Szilard: The Man Behind The Bomb..* New York, NY.: Skyhorse Publishing.

Legg, S. & Hutter., M., 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines,* Volume 17, p. 391-444.

Legg, S. & Hutter, M., 2007. A Collection of Definitions of Intelligence. *Frontiers in Artificial Intelligence and Applications,* Volume 157, pp. 17-24.

Leopald, C. A., 2014. Smart plants: Memory and communication without brains. *Plant Signalling & Behaviour,* 9(10).

Levanon, T., 2016. Thomas Reid and the Evolution of the idea of the SpeciousPresent. *History of Philosophy Quarterly,* 33(1), pp. 43-61 .

Levin, J., 2023. *Functionalism.* [Online]
Available at: The Stanford Encyclopedia of Philosophy, Edward N. Zalta & Uri Nodelman<https://plato.stanford.edu/archives/sum2023/entries/functionalism/>.
[Accessed 06 11 2024].

Lewis, D., 1980. Mad pain and Martian pain. In: *Readings in Philosophy of Psychology, Volume I.* Cambridge, MA.: Harvard University Press, p. 216-32.

LIFE Magazine, 1961. Some Szilardisms on War, Fame, Peace. *LIFE Vol.51 , No.9*, 1 September, p. 79.

Liljeholm, M., 2021. Agency and Goal-Directed Choice. *Current Opinion in Behavioral Sciences,* Volume 41, pp. 78-84.

Locke, J., 1688 . *An Essay on Human Understanding.* 1959 ed. New York: Dover.

Lowe, E. J., 2008. *Personal Agency: The Metaphysics of Mind and Action.* Oxford: Oxford University Press..

Lycan, W., 1987. *Consciousness.* Cambridge, Massachusetts: MIT Press..

Mallatt, J. et al., 2021. Debunking a myth: Plant Consciousness. *Protoplasma,* 258(3), p. 459-476..

Maravita, A., Spence, C. & Driver, J., 2003. Multisensory integration and the body schema: close to hand and within reach. *Current Biology,* 13(13), pp. 531-539.

Marcus, G., 2001. *The Algebraic Mind.* Cambridge, MA:: MIT Press..

Marques, H. .. & Holland, O., 2009. Architectures for functional imagination. *Neurocomputing,* Volume 72, pp. 743-759,.

McCann, H. J., 1998. *The Works of Agency: On Human Action, Will, and Freedom.* Ithaca: Cornell University Press..

McCarthy, J., 1987. Generality in artificial intelligence. *Communications of the ACM, Volume 30, Issue 12,* 30(12), pp. 1030 - 1035.

McCarthy, J., 2007. From here to human-level AI. *Artificial Intelligence,* 171(18), pp. 1174-1182.

McCarthy, J., Minsky, M., Rochester, N. & Shannon, C., 1955. A proposal for the dartmouth summer research project on artificial intelligence. *AI magazine,* 27(4), p. 12.

McClelland, J. L. & Ralph, M. A., 2015. Cognitive Neuroscience. *International Encyclopedia of the Social & Behavioral Sciences ,* Volume Second Edition, pp. 95-102.

McCulloch, W. & Pitts, W., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics; 7: ,* p. 115-133.

McDermott, D., 2001. *Mind and Mechanism.* Cambridge, MA: : MIT Press.

McDougall, W., 1923. *Outline of Psychology.* New York: Scribner.

Mele, A. R., 1992. *Springs of Action: Understanding Intentional Behavior, Oxford: Oxford University Press..* Oxford: Oxford University Press.

Mele, A. R. & Moser, P. K., 1994. Intentional Action. *Nous,* 28(1), pp. 39-68.

Mendelovici, A., 2013. Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies ,* 165(2), p. 421- 443.

Merleau-Ponty, M., 2012. Phénoménologie de la perception (1945). In: C. S. (. E. translation & D. L. (. E. translation), eds. *Phenomenology of Perception.* London: Routledge.

Merriam-Webster Dictionary, 2022. *Affordance.* [Online]
Available at: https://www.merriam-webster.com/dictionary/affordance.
[Accessed 8 July 2022].

Merriam-Webster Dictionary, 2022. *Bed.* [Online]
Available at: https://www.merriam-webster.com/dictionary/bed.
[Accessed 8 July 2022].

Merriam-Webster, 2020. *Sentience.* [Online]
Available at: https://www.merriam-webster.com/dictionary/sentience

Mollo, D. C., 2024. Intelligent Behaviour. *Erkenntnis ,* Volume 89, pp. 705-721.

Monett, D. & Lewis, C. W. P., 2017. Getting Clarity by Defining Artificial Intelligence—A Survey. In: *Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 44.* Berlin: Springer, pp. 212-214.

Moore, J. W., 2016. What Is the Sense of Agency and Why Does it Matter?. *Frontiers in Psychology ,* Volume 17.

Moravec, H., 1990. *Mind Children: The Future of Robot and Human Intelligence.* Cambridge, Massachusetts: Harvard University Press.

Moravec, H., 1998. When will computer hardware match the Human brain?. *Journal of Evolution and Technology.,* Volume 1, p. 10.

Morris, M. R. et al., 2024. Position: Levels of AGI for Operationalizing Progress on the Path to AGI. *Proceedings of the 41st International Conference on Machine Learning,* 235(Proceedings of Machine Learning Research).

Müller, V. C., 2024. Philosophy of AI: A Structured Overview. In: N. Smüha, ed. *Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence.* Cambridge: Cambridge University Press..

Nagel, T., 1974. What is it like to be a bat?. *Philosophical Review,* Volume 83, p. 435–456..

Newell, A., Shaw, J. C. & Simon, H. A., 1959. *Report on a General Problem-Solving Program.* s.l.:IFIP Congress.

Newell, A. & Simon, H. A., 1976. Computer science as empirical enquiry: Symbols and Search. *Communications of the ACM 19, 3:,* p. 113–126.

Newell, A. & Simon, H. A., 1976. Computer Science as Empirical Inquiry: Symbols and Search.. *Communications of the ACM,* 19(3), pp. 113-126..

Newen, A. & Vogeley, K., 2004. Self-representation: Searching for a neural signature of self-consciousness.. *Consciousness and Cognition 12(4):529-43,* 12(4), pp. 529-543.

Newton, I., 1687. *Philosophiae Naturalis Principia Mathematica.* London: s.n.

Norvig, P. & Russell, S. J., 2009. *Artificial Intelligence: A Modern Approach 3rd edition.* Saddle River, NJ: Prentice Hall.

O'Connor, T., 2000. *Persons and Causes: The Metaphysics of Free Will.* Oxford: Oxford University Press..

O'Gieblyn, M., 2021. *God, Human, Animal, Machine: Technology, Metaphor, and the Search for Meaning.* New York: Anchor Books.

OpenAI, 2019. *OpenAI Five defeats Dota 2 world champions.* [Online]
Available at: https://openai.com/index/openai-five-defeats-dota-2-world-champions/
[Accessed 8 11 2024].

Orseau, L., McGill, S. M. & Legg, S., 2018. Agents and Devices: A Relative Definition of Agency. *ArXiv,* Volume abs/1805.12387.

Oxford Languages, 2023. *automaton.* [Online]
Available at: https://www.google.com/search?sca_esv=583590671&rlz=1C1CHBF_en-GBGB836GB836&q=automatons&si=ALGXSlbK6dNKc3P-z0hratVoTzWI5ab5y8Y4GcQiYKULGqs4TJ6wOzQBs05LbBgPmoigkMQP3__9uamcwuSq4GDD1dUYXv9DGpkTdogoTe1R76zZbkVY2AQ%3D&expnd=1&sa=X&ved=2ahUKEwi066yAqs2CAxW3QE

Parker, B., 2011. The Tide Predictions for D-Day. *Physics Today 64(9).*

Peper, F., 2017. The End of Moore's Law: Opportunities for Natural Computing?. *New Generation Computing,* Volume 35, p. 253-269.

Pezzulo, G. & Castelfranchi, C., 2009. Intentional action: from anticipation to goal-directed behavior. *Psychological Research PRPF ,* Volume 73, p. 437-440.

Pinker, S., 2005. So How Does the Mind Work?. *Mind and Language, 20: ,* p. 1-24.

Pinotsis, D. A., 2007. The Antikythera mechanism: who was its creator and what was its use and purpose?. *Astronomical and Astrophysical Transactions,* 26(4-5), pp. 211-226.

Prinz, W., 2019. Import Theory: The Social Making of Consciousness. *Journal of Consciousness Studies ,* 26(3-4), pp. 112-130.

Proudfoot, D., 2017. Child machines. In: Online, ed. *The Turing Guide.* Oxford: Oxford Academic, p. 315-325.

Putnam, H., 1960. Minds and Machines. In: *Dimensions of Minds..* New York: New York University Press, pp. 138-164.

Putnam, H., 1983. *Realism and Reason: Philosophical Papers, vol. 3..* Cambridge: : Cambridge University Press..

Putnam, H., 1988. *Reality and representation.* Cambridge, MA:: MIT Press..

Pylyshyn, Z. W., 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge: MIT Press.

Rathi, N. et al., 2023. Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. *ACM Computing Surveys,* 55(12), pp. 1-49.

Rescorla, M., 2015. *The Computational Theory of Mind.* [Online]
Available at: https://plato.stanford.edu/entries/computational-mind/
[Accessed January 2022].

Reuters, 2024. *Microsoft, OpenAI plan $100 billion data-center project, media report says.* [Online]
Available at: https://www.reuters.com/technology/microsoft-openai-planning-100-billion-data-center-project-information-reports-2024-03-29/
[Accessed 04 08 2024].

Rhodes, R., 1986. *The Making of the Atomic Bomb.* New York, NY.: Simon and Schuster.

Rumelhart, D. E. & McClelland, J. L., 1986. PDP Models and General Issues in Cognitive Science. In: *Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1.* Cambridge, MA: : MIT Press, pp. 110-146.

Russell, S. & Norvig, P., 2003. *Artificial Intelligence: A Modern Approach.* New Jersey: Prentice Hall.

Sartre, J.-P., 1956. L'Etre et le néant 1943. In: H. E. B. (. E. t. S. R. (. E. translation, ed. *Being and Nothingness.* New York: Philosophical Library.

Sayre, K., 1986. Intentionality and Information Processing: An Alternative Model for Cognitive Science. *Behavioral and Brain Sciences ,* 9(1), pp. 121-138..

Schank, R. C. & Abelson, R. P., 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures..* Hillsdale, NJ: L. Erlbaum.

Schlosser, M., 2019. *Agency.* [Online]
Available at: https://plato.stanford.edu/archives/win2019/entries/agency/
[Accessed November 2022].

Schrittwieser, J. et al., 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature ,* Volume 588, p. 604-609 .

Schultz, W., 2015. Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews,* 95(3), p. 853-951.

Searle, J., 1980. Minds, brains, and programs.. *Behavioral and Brain Sciences,* 3(3), pp. 417-457.

Searle, J., 1984. *Minds, Brains and Science.* Cambridge, Mass.:: Harvard University Press..

Searle, J., 1990. Is the Brain a Digital Computer?. *Proceedings and Addresses of the American Philosophical Association,* 64(3), p. 21-37.

Shanahan, M., 2015. *The Technological Singularity.* s.l.:MIT Press.

Shannon, C. E., 1938. A Symbolic Analysis of Relay and Switching Circuits. *Transactions of the American Institute of Electrical Engineers,* 57(12), pp. 713-723.

Shettleworth, S. J., 2010. *Cognition, evolution, and behavior.* 2nd ed. New York: Oxford University Press.

Shoemaker, S., 1990. The Intentional Stance by Daniel Dennett. *The Journal of Philosophy,* 87(4), pp. 212-216.

Siewert, C., 1998. *The Significance of Consciousness..* Princeton, NJ: Princeton University Press..

Sloman, A., 1978. *The Computer Revolution in Philosophy.* Hassocks:: The Harvester Press.

Sloman, A., 2021. *The Self - A Bogus Concept? Yes and No.,* Birmingham UK: s.n.

Smolensky, P., 1988. On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences,* Volume 11, p. 1-74..

Sonni, J. & Goodman, R., 2018. *A Mind at play.* Gloucestershire: Amberley Publishing.

Spearman, C., 1904. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology,* 15(2), pp. 201-292.

Stampe, D. W., 1977. Toward a Causal Theory of Linguistic Representation. In: *Midwest Studies in Philosophy, vol. 2,.* Minneapolis: University of Minnesota Press, p. 42-63.

Stich, S., 1983. *From Folk Psychology to Cognitive Science.* Cambridge, MA: : MIT Press..

Strawson, G., 1994. *Mental Reality, Cambridge, MA: MIT Press.* Cambridge, MA: MIT Press.

Strawson, G., 1997. The Self. *Journal of Consciousness Studies,* 4(5), pp. 405-428.

Strawson, G., 1999. The Self and the SESMET. In: *Models of the Self, Imprint Academic.* Exeter, UK.: Imprint Academic, pp. 483-518.

Strawson, G., 2011. The Minimal Subject. In: S. Gallagher, ed. *The Oxford Handbook of THE SELF.* New York: Oxford University Press, pp. 253-278.

Tallon, A., 1997. *Head and heart: affection, cognition, volition as triune consciousness.* New York: Fordham University Press.

Taylor, R., 1966. *Action and Purpose.* Englewood Cliffs: Prentice-Hall.

The AlphaStar team, 2019. *AlphaStar: Mastering the real-time strategy game StarCraft.* [Online]
Available at: https://deepmind.google/discover/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii/
[Accessed 8 11 2024].

Thicknesse, P., 1784. *The speaking figure, and the automaton chess-player, exposed and detected..* London: Wellcome Collection.

Thomas, E., 2023. The Specious Present in English Philosophy 1749-1785: Theories and Experiments in Hartley, Priestley, Tucker, and Watson. *Philosophers' Imprint ,* 23(7).

Thomson, G. H., 1916. A hierarchy without a general factor. *British Journal of Psychology,* 8(3), p. 271.

Trewavas, A., 2016. Intelligence, cognition, and language of green plants. *Frontiers in Psychology,* Volume 7, p. 588.

Tsakiris, M., 2016. The multisensory basis of the self: From body to identity to others. *The Quarterly Journal of Experimental Psychology ,* 70(4), pp. 597-609.

Turing, A., 1948 (1992). Intelligent Machinery. In: *Mechanical Intelligence, Collected Works.* Amsterdam: North Holland, p. 107-127.

Turing, A. M., 1936. *On computable numbers with an application to the decision problem,* London: Proceedings of the London Mathematical Society.

Van Gulick, R., 2022. *Consciousness.* [Online]
Available at: https://plato.stanford.edu/archives/win2022/entries/consciousness/>
[Accessed 24 08 2024].

Vaughan, R. T. & Zuluaga, M., 2006. *Use Your Illusion: Sensorimotor Self-simulation Allows Complex Agents to Plan with Incomplete Self-knowledge.* Rome, Italy, From Animals to Animats: 9th International Conference on Simulation of Adaptive Behavior.

Vinge, V., 1993. The Coming Technological Singularity: How to Survive in the Post-Human Era. In: *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace.* s.l.:NASA Publication, p. 11-22.

Ward, J. & Meijer, P., 2010. Visual experiences in the blind induced by an auditory sensory substitution device. *Consciousness and Cognition,* 19(1), pp. 492-500.

Weinstein, V. K. et al., 2024. An Internal Model Principle For Robots. *arXiv:2406.11237v1,* Volume 1.

Wells, H. G., 1914. *The World Set Free: A Story of Mankind.* London: Macmillan & Co.

Willis, R., 1821. *An attempt to analyse the automaton chess player, of Mr. de Kempelen.* London: Booth.

Winfield, A. F., 2014. Robots with Internal Models: A Route to Self-Aware and Hence Safer Robots. In: J. Pitt, ed. *The Computer after Me: Awareness and Self-Awareness in Autonomic Systems.* s.l.:World Scientific, pp. 237-252.

Wonham, W., 2018. *The Internal Model Principle of Control Theory.* [Online]
Available at: https://www.control.utoronto.ca/~wonham/W.M.Wonham_IMP_20180617.pdf
[Accessed 20 Ocotber 2024].

Xabier E. Barandiaran, E. D. P. M. R., 2009. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior,* 17(5), p. 367-386.

Yang, J. et al., 2020. General learning ability in perceptual learning. *Proceedings of the National Academy of Sciences,* 117(32), pp. 19092-19100.

Zahavi, D., 2002. First-person thoughts and embodied self-awareness: Some reflections on the relation between recent analytical philosophy and phenomenology,”. *Phenomenology and the Cognitive Sciences,* 1(1), p. 7-26..