An Eye for an 'I':
Investigating the Relationship between Embodiment and Agency and the Possibilities for Artificial Intelligence Models

MSc Thesis *(Afstudeerscriptie)*

written by

Kira G. B. Miller

under the supervision of **Dr Tom Schoonen**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
August 25, 2025	Dr Tom Schoonen
	Dr Karolina Krzyzanowska
	Dr Giorgio Sbardolini

Institute for Logic, Language and Computation

Abstract

This thesis examines the relationship between embodiment and agency with specific reference to Artificial Intelligence (AI) models. It argues that a sense of embodiment is a necessary condition of agency because agents must be able to sense and manipulate the external world that they inhabit. I argue that although some AI models are low-level agents, high-level embodied agency has the potential to help AI models overcome their current limitations by allowing them to learn from and ground meaning in first-hand experience. I explore existing robotic technology and agentive machine learning architectures to make the case that embodied artificial agents are possible and, if achieved, have the potential to be significantly more powerful, efficient, and versatile than current AI models.

Contents

Chapter 0

Introduction

Agency has often been cited as one of the key human capacities which differentiates us from automata [Hue21] [DeL11] [Lis21]. Humans (and a number of other intelligent beings) are capable of intentionally acting in a particular way in pursuit of particular goals. Beyond the capacity that most organisms possess to move towards pleasure (food, sunlight, water, etc) and away from pain (predators, danger, poisons, etc), some organisms are able to internally represent an expectation of how a particular action might transpire, allowing them to choose the action which might most align with their goals. For example, if I am faced with the choice of what activity to do with a group of friends, I can envision what the experience of each activity might be like and make a selection based my preferences (for certain activities, foods, locations, etc), short term goals (to see a particular show, exercise, have a conversation, etc), and my long term goals (to forge connections with my friends, explore new places, develop my social skills, etc). In addition to simple, immediate responses to stimuli, I am able to make considered decisions based on my environment, experiences, and the other agents around me that align with my values, desires, and goals. Some other organisms possess the same ability at a lower level; they may pursue simpler, shorter-term goals using less detailed outcome expectations and less understanding of the external world. While the precise definition of agency is the subject of extensive debate amongst philosophers [Ari99] [Hum07] [Dav01] [Ban89], it is clear that a fundamental capability of intelligent organisms is the ability to act intentionally for a particular purpose.

Thus far, artificial intelligences (AI), largely do not possess high levels of agency. To give a broad definition, 'AI' refers to the general ability of computers to perform tasks commonly associated with intelligent beings. AI programs have, in the last several decades, managed to attain or even exceed the performance level of human experts in a range of specific tasks. For example, in March 2016 a Go-playing program developed by DeepMind called AlphaGo defeated top (human) Go player Lee Sedol in a five-game match, winning four of the five games [Wan+16]. Further still, AI systems have found applications in early cancer diagnosis, optimising search engines, voice, or handwriting detection, and chatbots. While the most advanced AI models are capable of solving some highly complex problems and calculate optimal actions for maximising a reward (such as points in a game), the level of agency they possess is limited. They are limited by how much they know about the world in which they exist, the complexity and accuracy of the predictions they can make and how much their actions are able to make changes in the world.

Further still, there are a number of limitations that the AI models that currently exist have yet to overcome. Machine Learning (ML) models, which are AI models with the capability to learn autonomously, require a great deal more data than seemingly required by humans in order to learn insights, recognise patterns, and apply that learning to make increasingly better decisions. For example, large language models (LLMs) are ML algorithms that uses massive amounts of training data to understand and predict text. They require a great deal more data to learn how to properly use language than humans do [CWF24]. ML models also struggle to transfer learned solutions to similar problems, understand causality, and struggle to explain their decision-making processes, to name a few further issues. In addition, some artificial intelligence research attempts to pursue the creation of artificial general intelligence (AGI), sometimes called strong AI. An AGI would be a machine with intellectual abilities indistinguishable (or perhaps exceeding) that of a human being. One of the key issues that researchers face in pursuit of this goal is generalisability. While AI models can perform exceedingly well as highly specialised tasks, they perform poorly when solving a broader range of problems. Thus, paralleling the human ability to perform tasks over multiple domains, adapt knowledge from past experience, self-reflect, and generally understand the external world as a whole, is out of reach for current AI systems.

This thesis aims to investigate whether the limitations of current AI models may be overcome by allowing these models to achieve higher levels of agency, and how we may go about doing so.

One way that researchers have attempted to improve AI models and bring them closer to replicating human intelligence is by developing models that are more similar to human brains. Neural Networks, for example, are computational models that mimic networks of neurons, similar to those that exist within the human brain. They consist of interconnected nodes that process data, and are used to approximate non-linear functions. These nodes are usually arranged into layers. The term 'Deep Neural Network' (DNN) typically refers to a neural network with more than three layers.

For this reason, it is worthwhile to turn to research into human development to gain some insight into how humans are able to achieve the range of cognitive abilities that we possess. The term 'cognitive ability' will be used to refer to the skills involved in performing tasks associated with perception, learning, memory, understanding, reasoning, judgement, intuition, and language [Nei+96]. According to psychologists such as Jean Piaget [Gau09], infants develop knowledge and understanding through directly experiencing and manipulating objects. As development continues, children are able to learn from others, through the imagination of future and possible scenarios, and eventually through more advanced logical reasoning.

Abilities that aid in the development of human cognitive capabilities, such as sensing and manipulating external objects, imaging possible scenarios, and testing hypotheses, are also the abilities that make up agency. The more developed these skills are, the higher the level of agency a being might be able to achieve.

In Chapter 1 of this thesis, it will be argued that a higher level of agency has the potential allow AI models to overcome several of their current limitations. If models are able to make better predictions about short and long term outcomes in addition to having a high level of control over the external world, then they might be able to perform more detailed investigations and experiments, develop a broader understanding of the world, and utilise external objects to aid in the execution of cognitive

tasks.

In Chapter 2, it will be argued that a sense of embodiment is a requirement for agency. As humans, we understand ourselves as beings within the world. The way that we experience, think, make decisions, and act is inseparable from our physical extension in a physical world. I am able to experience the world through the use and adjustment of my sense organs. I am able to make changes in the world by using my body to speak, write, manipulate objects, explore, etc. My ability to sense and manipulate the world is what allows me to develop increasingly accurate predictions for the outcomes of my actions, to refine my ability to determine the actions that best align with my goals, and to improve my motor function such that I have a greater level of control over my actions. While it is possible to be an agent over a limited world (for example, the world of a simple game), the level of agency that the agent can achieve is similarly limited. It will also be argued that embodied agency is what allows for a high level of agency.

In Chapter 3, it will be argued that there are model architectures and physical components that already exist that demonstrate both that higher levels of agency in AI models are possible to achieve, and that making such changes could in fact help to lessen, or even entirely remove, the limitations faced by current models. As it was shown in Chapter 2 that a sense of embodiment is a requirement for high levels of agency, it will be argued that providing models with access to physical components which meet the criteria for embodiment (as well as a model architecture that allows them to effectively utilise these components) is how higher levels of agency can be achieved. For example, providing real-time, multimodal data, such as a continuous visual and audio stream has the potential to greatly reduce the data requirement for a number of problems, such as image categorisation, as it allows for the understanding of how objects move and change over time as well as filling in any gaps from one modality with the other. In addition, there are models which allow for the adjustment and "pointing" of sensing devices towards areas of interest, the goal-based determination of action procedures, and learning from outcomes in order to update and refine future actions. It will be argued that a combination of the architectures and physical components that already exist could be used to provide models with a level of agency similar to that of humans, and that embodiment and agency may possibly push AI research further towards the development of an AGI.

Overall, I propose to demonstrate the significance of agency in the development of high-level cognitive abilities, which themselves aid in reaching higher levels of agency. I will argue that embodiment is fundamentally necessary for a high level of agency, and that providing AI models with the ability to directly sense and manipulate the external world is therefore essential for the development of high-level artificial agents. Such agents, with their broader understanding of the external world and ability to explore and experiment within it, have to potential to be able to solve highly complex problems over a range of domains, pursue long and short-term goals, reflect upon their decision-making processes, and learn from a great deal less data than models currently require.

Chapter 1

This chapter aims to understand the ways in which high levels of agency might be a requirement for certain mental capabilities found in humans. This question is particularly relevant for understanding the possibilities for artificial intelligence (AI) models. AI models such as Large Language Models (LLMs) are already capable of a great deal, especially when it comes to specific tasks and synthesising information from across the Internet in a conversational style. However, AI models are still limited in a number of ways. This chapter aims to explore what these limitations are, and whether some of these limitations may be overcome by instilling AI models with stronger agency. I will begin by giving an overview of the ways in which existing AI models are limited, followed by a brief discussion of how this compares to human abilities and development in order to identify the areas in which agency may play a role. I will end this chapter by arguing for the areas that I believe will benefit from developing agency in AI models.

1.1 Current Technological Limitations

As previously stated, current AI models have highly advanced abilities such as language skills and complex reasoning and can generate new images or future scenarios. But it is what they are not yet capable of doing that might give us a greater insight into how agency fits into the cognitive network. By understanding the current limitations of AI models, we can begin to understand where advancing agency might be of value.

1.1.1 Buckner's Limitations

We begin with some of the limitations of machine learning (ML) models identified by Cameron J Buckner in his 2023 book *From Deep Learning to Rational Machines: Perception, Memory, Imagination, Attention, and Social Cognition* [Buc23]. Buckner specifically discusses ML models as opposed to AI models in general, as non-ML models are generally able to handle much less complex problems because they are largely rule-based and follow explicit decision-making patterns[Inf25]. These are:

- **Opacity:** ML models often lack the ability to explain how they reached a certain decision. For example, Deep Neural Networks (DNNs) have millions of parameters, so it can be nearly impossible for humans to interpret which factors were most significant in leading to certain decisions. There have been some developments in this field since Buckner's writing, with more LLMs having the ability to cite directly the sources that they utilise in their answers. For example, Microsoft Copilot uses a pre-processed library of documents to find relevant chunks

of text to use in its answers [Lin+24]. That said, Copilot still lacks the ability to explain why specific sources are chosen and how answers are compiled on a case-by-case basis. In general, opacity remains an issue for AI models. There are a great number of proposed solutions to this issue, and we will examine whether agency might be part of these solutions or perhaps a natural consequence of them.

- **Training Efficiency:** What has become known as 'The Data Efficiency Crisis in Machine Learning' refers to the concern that ML models currently require very large amounts of data to perform well. Models require much more data than humans do for equivalent levels of performance [Ada21]. For example, AlphaGo's networks were trained on more than 160,000 stored games, before it then played millions of games against iteratively stronger versions of itself. Meanwhile, its human counterpart, Lee Sedol, would have played no more than 50,000 matches in his lifetime [Buc23](pg.65). Some of the proposed solutions to this problem include the recognition that humans receive data very differently from most ML models. For example, as babies learn to interpret the physical world, they receive a continuous stream of multisensory data and are able to direct their sense organs towards particular objects [MN23]. This hints towards agency being a significant factor in resolving the data efficiency crisis, and we will return to this later in the chapter.

- **Adversarial Attacks:** ML models can often be misled by what are known as 'adversarial examples'. These are unusual stimuli that are specifically designed to fool or mislead a model [GSS15]. For example, an image classification model might be fed an image that is altered in a way that is imperceptible to humans but that leads to a dramatic misclassification by the model. This issue again points to the fact that the way these models receive and process data is very different from the way humans do. Altering this may, therefore, allow models to better overcome these kinds of adversarial attack, or perhaps develop the ability to identify and ignore the alterations to the stimuli.

- **Reward Hacking:** Many models come up with solutions to problems that are entirely different from the intended outcome. Many describe this as models lacking "common sense" that a human might bring to a task. For example, OpenAI attempted to teach a DNN to play a boat racing game called *Coast Runners*. Instead of actually competing in races, the model endlessly turned the boat in circles in order to receive "turbo" bonuses and continually boost it's score [CA16]. One of the solutions to these kinds of problem is simply to provide the model with a broader understanding of the context of the problem and a more detailed understanding of the goal (e.g. "win as many races as possible", as opposed to "get as many points as possible"). At first blush, this does not seem like an issue for which agency is a requirement to solving, as increased prompt specificity and greater understanding of the context of the task can be achieved without it. However, the ability to understand the broader context of the task could help to prevent models from reward hacking.

- **Transfer Learning:** ML models often struggle with the ability to transfer the solutions and abilities learned in the context of one problem to another. A model trained to play a particular game might struggle to adapt to even small changes in the game environment. For example, when a ML model was taught to play the Atari game *Breakout*, a game in which the player

moves a virtual paddle horizontally to bounce a ball towards targets, it struggled to adapt to playing the game when the paddle was moved up on the Y-axis, while humans took very little time to adapt [GG18]. This points to the idea that ML models do not learn the same kinds of representations of their environment as humans do. The ways in which they learn about and interact with the world are very different from the ways that we do. While the ability to act intentionally may not factor into transfer learning particularly, we will discuss the consequences a model having a limited and almost unchanging set of objects over which it has direct control. When interaction with the world always involve some kind of familiar action, such as the movement of an object over which one has always had direct control and through which one has always received sense feedback, it might perhaps be easier to adapt the steps required to complete one task for another, similar task. In the *Breakout* case, human players were still performing the similar basic physical actions (hand and eye movements), whereas the change for the model was akin to controlling one's own hand from a meter away from where it usually is. It may be that being able to sense with and control a consistent set of objects that will be used across all tasks is the reason why humans are able to adapt more easily to new scenarios: the tools remain the same.

- **Compositionality and Systematicity:** Another issue that has been raised with regards to AI models is whether they are capable of compositionality and systematicity. This is a capability of human thought where, based on a limited number of mental representations, a potentially infinite number of new thoughts can be produced. It could be argued that, in the age of generative AI, that this is already a capability that machine learning models possess. Whether or not this is "truly" the case is an ongoing debate which is beyond the scope of this paper. It will be enough for our purposes to acknowledge that AI models are able to generate original (in that they are not exact copies of input data) and coherent texts, images, and even videos (take GPT-4 and Dall-E, for example).

- **Causal Understanding:** Perhaps the cognitive gap that most prominently comes to mind when we consider agency is that of causal understanding. Humans are fairly adept at determining causal relationships, as opposed to mere correlations. This kind of thinking is vital for making predictions about the result of certain actions, counterfactual thinking, and transferring learning. Judea Pearl [Pea09] argues that all AI models have been able to achieve so far is the recognition of patterns or correlations, but that they cannot predict the effect of actions or reason with counterfactuals. Buckner describes figuring out how to best model causal understating as "perhaps the most important unsolved problem in the field and the one most likely to require a substantially new kind of structure in standard DNN architectures to overcome"(pg 74). In the next part of this chapter, I will make the case that if a model were to be given agency – in particular, the ability to interact with the world, observe the consequences of those actions, and adapt future actions accordingly – some form of causal understanding must arise. We will also need to examine whether causal understanding is something that arises from agency, or if it is something that is fundamental to agency itself. If it is the latter, then we must delve deeper into exactly which skills lead causal thinking to arise.

- **Emotions and "Care":** Finally, we turn to a limitation of ML models that is mentioned incredibly often discussions of ML models – feelings. Models do not have emotions. A model

has no reason to "care" one way or the other about a particular thing. A model does not have a reason to pursue any particular goal other than that we have told it to. The exact nature of emotions and whether it is cognitively important for AI models to possess them is far beyond the scope of this thesis, but we will discuss the implications of how a model might be able to not only recognise the emotions of other agents and demonstrate empathy for them, but also perform more complex social and moral reasoning.

1.1.2 Further Limitations

The limitations identified by Buckner are certainly not all modern AI models face. Buckner's writing pertains specifically machine learning, whereas our discussion looks more broadly at AI models as a whole. There are some further limitations that I have identified and which are important to discuss.

- **Metacognition:** AI models largely lack the ability to think about thinking. They are currently not able to monitor their own thoughts or recognise mental states (if they have any) as they are taking place. Additionally, AI models are unable to self-critique. They rely on external stimuli to determine when they have made a mistake, rather than being able to check and evaluate their own responses. Some models are able to give estimates of the confidence they have in their predictions, but this is not the same as self-reflection and evaluating their own decision-making processes. If, however, a model is able to generate an expectation of an outcome and compare it against the actual consequences of a particular action, then this may at least give rise to the kind of metacognition in which an agent understands that they had a particular mental state (an intention) and a thought (an expectation).
- **Hallucination:** Large Language Models (LLMs) are prone to what is known as hallucination [Hua+25]. This occurs when an LLM generates plausible, yet false information. Given that they largely generate text on a probability basis, responses are not often checked for accuracy and truthfulness. Some amount of goal-directedness towards truthfulness would push these models towards producing factual information; however, it seems here that what is really required is a system of fact-checking and a requirement for providing sources. Agency is likely not a requirement for overcoming this issue.

1.2 Are Some Models Already Agents?

Having established a list of key limitations of current AI models, it is now time to delve deeper into what I mean by a 'higher level' of agency. To begin with a simple definition: in order to have agency, a being needs to be able to act (interact with the external world) intentionally (with an expectation of desired outcome) and be able to compare the outcome of that action to the expected outcome. In the next chapter, I will argue in depth for this definition. In this chapter, I will use this intuitive understanding of agency to argue for how it is instrumental in the development of certain cognitive abilities.

To make the case for how agency may help overcome the limitations of modern AI models, I wish to briefly mention how agency is involved in the development of the mental capabilities possessed by humans. Understanding the role that agency plays will help point to where it may aid AI models and help them progress towards more advanced skills. Animals have largely evolved agency and cognitive

capabilities alongside one another. At the most basic level, an organism with a low level of agency might only be able to act in pursuit of immediate gratification. What makes this organism capable of action, as opposed to reflex, is some kind of mental state that represents a desired outcome that led to the particular movement. A being with a high level of agency is capable of pursuing much more complex goals over a range of time scales and generating detailed expectations of action consequences. They have a much more in-depth understanding of the external world and the ability to act with a great deal more precision and control. In the next chapter, I will argue that higher levels of agency are achieved when an agent is able to receive detailed sense data about the external world and the ability to manipulate the external world.

In this section, I want to evaluate whether or not current models are agents and to what extent. I will argue that there are some models that could be considered agents, but only weakly. This will shed light on what would need to be changed in order for a model to be considered a higher-level agent and why this might lead to enhanced abilities.

1.2.1 Intentionality

As I will argue, agency requires the ability to manipulate the external world in some way, meaning that actions are external. I use the word "external" as opposed to "physical", as I will allow for scenarios in which an agent acts within a simulated world. So, whatever world the agent can sense and manipulate need only be external to the agent (it cannot be generated by the being itself), but may be simulated by some other device. Doing this will also avoid begging the question when it comes to embodiment, as physicality has not been assumed.

Let's look at some AI models in order to evaluate whether or not they are agents, and, if so, to what extent. Take, for example, an LLM like GPT-4. Is its output in conversation with a human user considered an act? One might compare this type of output to human speech. There is no doubt that when a human speaks, they are performing an action. But the difference here is that the human in this scenario is still employing physical movement of a body to speak, which is not the case for the LLM. When a human and an LLM engage in 'conversation' (for lack of a better term), the human makes physical movements (such as typing or speaking) in order to input their responses to the LLM. The LLM, however, does not, though we can argue that the external world does change when the output is displayed. The difference here is that the LLM is not capable of a physical action.

A comparison we might make is that the output of an LLM is like if a human were capable of telepathy. To take this thought experiment further, suppose this telepathic human had never received any form of sense data other than telepathic responses from other humans. Their only method of "acting" is sending these telepathic messages, and their only method of "sensing" is receiving them. If we would call such a person an agent, then there is scope for suggesting that LLMs, as they currently exist, are agents.

It is my view that even if we concede that this type of non-embodied communication counts as action (after all, it has brought about a physical effect in the world), this type of LLM is not an agent, because it is not an *intentional* action. The ability to act intentionally is part of the definition of agency. The LLM does not intend a specific effect when it "speaks", it simply produces the response predicted by its specific architectural makeup. It might be able to, if asked, come up with a prediction of the response

it might receive, but it does not alter its outputs in order to achieve a specific response because it lacks goal-directedness. LLMs generate responses by predicting the most plausible next word or sentence based on the input and training data. It does not have internal motivations or desires or plan its responses in order to reach a particular goal.

It might be argued in response that the telepathic human in the earlier thought experiment does not face these issues. There is no reason why this person may not wish to elicit a specific type of response from their interlocutor, and they may alter their telepathic messages in order to pursue this goal. So, would instilling some form of goal-directedness in LLMs be enough for them to be considered agents? To use another example from AI, a game-playing model such as Alpha-Go acts with the intention of winning the game / scoring the most points.

1.2.2 Expectation

In addition to intention, as discussed in the definition given at the beginning of this section, the agent must have some sort of expected outcome for the action which they are about to undertake, coupled with a desire to achieve said outcome. In their 2003 paper "The Self as an Embodied Agent", Chris Dobbyn and Susan Stuart [DS03] consider requirements for a self-aware agent. They argue that the goals of an agent should be self-directed – the ultimate end is to bring about some internal state of the agent (for our most basic agent this might be pleasure), rather than a state in the external world. They outline four specific internal states required for this process: first, the agent is aware of its current internal state, secondly it has a conception of an internal state that it would like to achieve, thirdly it executes a plan to achieve this goal, and finally a new internal state is formed. Dobbyn and Stuart note here that more complex goals may involve a more complex version of this process, where intermediate subgoals must first be achieved.

This description aligns with my own discussion, as I will take the plan for achieving the desired internal state (the intention) to be the same as an expectation: the agent carries out a particular plan because it expects the outcome to lead to the desired internal state. For example, if I am hungry and I move towards the fridge to retrieve some food, I have the expectation that I will experience the sensations associated with this process (moving my body, feeling and seeing my hands open the fridge, tasting the food). These are the sensations I expect to experience in order to achieve my goal of feeling satiated, and, because I am receiving continual sense data, if my expectations are not met, I can adjust my plan as I am carrying it out. The complexity of my expectation is correlated with the amount of sense data I have access to and my ability to interpret said data.

AI models are programmed to be goal-directed in the sense that they are designed to carry out a set of tasks, but they might set expectations in different or less detailed way that humans do. An AI model might be able to set expectations about the feedback it might receive, if it is able to internally simulate the feedback it expects to receive. A model must therefore be working with some set of (however minimal) assumptions with regards to the laws that govern their world, e.g. "action X will result in the receipt of data Y". If the expected sensation does not occur, then these rules are updated.

Let's return to the example of a game-playing model such as AlphaGo. In addition to its goal to win/score more points in the game, these models possess some amount of expectation of how the board/game space will appear after they make a particular move and the events that might take place

within the game after this. This is how it determines which move it will make. Therefore, is model an agent?

At first glance, our instinct might be to answer 'no'. However, the telepathic human and Alpha go both meet the definition of agency as outlined in the start of this chapter. They can act intentionally, with (potentially) some amount of expectation of outcome. An agent need not be able to interact with the physical world, simply an *external* one. So by our definition there is no reason why, for example, the world of Go, although extremely limited, does count as a world that AlphaGo is sensing and manipulating. Thus, we conclude that this model *is* an agent. Even if the actions it is capable of are physical in the sense of physically moving pieces or displaying moves on a screen, as far as the model can determine, the act exists purely within the world of the Game. In this way, AlphaGo, and goal-directed models like it, have limited agency, a conclusion which resonates with our understanding of agency as a spectrum, rather than something a being does or does not have. They are limited because the world they can sense and manipulate is limited. Models without goal-directedness, such as LLMs and other models whose outputs are based in probability, are not agents.

These conclusions lead us to a slight paradigm shift in how we understand the relationship between agency and cognitive capabilities. Firstly, some cognitive capabilities may require not just agency, but *embodied* agency. If sensing and interaction is limited to a non-physical external world, then it makes sense that cognitive capabilities might be limited compared to creatures, such as humans, that have the ability to sense and interact with the physical world. When agents are embodied, they are able to sense and manipulate the physical world that we inhabit. Agency is a spectrum that is determined by how much the agent can sense about the world and how much they are able to interact with it. As such, being able to solve complex problems about, and react in real time to, the physical world requires a higher level of agency than AI models currently have, which is achieved via embodiment.

1.2.3 A Spectrum of Senses

This leads us to the discussion of the ways in which an embodied agent might sense and manipulate the physical world.

The level of agency is made higher not only by the amount of information that a being is able to receive (e.g. being able to see and hear), but also how well they are able to make sense of this data. At a more basic level, an organism might be an agent in that they act with the intention of avoiding pain and seeking pleasure, and the sense information they receive is as simple as informing the organism which of these they are experiencing (these senses are external in the sense that they come from an external source). They are then able to alter their actions based on the past experience of which caused them pleasure and which caused them pain. At a more complex level, a human is able to interpret a great deal about the world based on visual, auditory, taste, smell, and touch data (as well as other senses such as proprioception). Being able to interpret this data, a skill we develop because we are able to explore and interact with the world and direct our senses toward the consequences of our actions, allows us to pursue more complex long-term goals.

The comparison of my expectations with the actual outcome is what allows me to distinguish myself as a being in the world: there are objects (my body) over which I have direct control, and there are certain sensations that are caused by the manipulation of those objects. This is yet another capability

which AI models do not (yet) have). The level of sensation and agency that humans have allows us to distinguish the boundaries between our own bodies and the rest of the physical world.

This points us towards the idea, which we will expand on in the next section, that cognitive development is greatly aided by agency, which itself is developed through the employment of some cognitive abilities. My agency allows me to explore the world and uncover more information about it. This information allows me to build more detailed and more accurate predictions for the outcomes of my actions, allowing me to produce more precise motor signals and develop more control over the world around me which, in turn, allows me to solve more complex problems. An model that is given more multimodal data about the world in which it acts and given greater resources to interact with that world is likely to understand the world more deeply. However, this simply describes (perceived) embodiment. The missing piece of the puzzle is intentionality: acting in pursuit of a particular goal and generating an expectation of the outcome.

1.2.4 Overview

In this section, we have explored the ways in which current AI model can sometimes be considered agents, and how they are currently limited in their agency. At minimum, it must be able to act within, and receive data from, an external world. These acts must be intentional, carried out with the expectation of receiving particular data and achieving a desired internal state. Some models, such as LLMs, do not act intentionally, so they are not agents. Some models, such as game-playing models, are limited agents because they act with intention and generate expectations for the outcomes of their actions. Agents who receive data in greater volumes and a broader range of modalities have to potential to be able to generate more complex outcome expectations and pursue a more complex range of subgoals. Expectations are what allow agents to distinguish themselves from the external world. A higher level of agency is achieved when agents have multiple methods of interacting with their external world and more control over how they do so.

Most AI models only have the ability to work with very limited virtual worlds, such as AlphaGo, which only works within the world of the game. When an agent is embodied, it has access to information about and the ability to manipulate the physical world. A model like AlphaGo is not embodied because it only has agency over the Go board. Even if this board were to be physical, the rest of the physical world, as far as it is concerned, does not exist. Embodiment allows for the possibility of receiving sense data from the whole physical world.

A high level of embodied agency allows the possibility of solving the kind of complex and far-reaching problems that humans can solve. I leave open the possibility of the model acting entirely within a virtual world that simulates the physical world, as opposed to being embodied in the physical world, but note that achieving a sophisticated simulation such as to allow for a high level of agency would be a very difficult endeavour to achieve, from a technological standpoint [Vid14].

1.3 Filling the Gaps

We can finally turn to understanding whether any of the cognitive limitations facing current AI models can be resolved by implementing a high level of embodied agency. Understanding this means

understanding what happens when they possess, at minimum, the ability to generate expectations. In addition, when they are given the ability to sense and interact directly with the physical world.

1.3.1 Data Efficiency

In his 1998 paper "Embodiment and the Philosophy of Mind", Andy Clark [Cla98] makes the argument that there are a great deal of "computational economies" (pg. 29) that are afforded by action within the real world. For example, a deictic pointer is a bodily orientation that "points" to other data structures. This allows us to focus on a particular part of a visual scene and retrieve more detailed information. In addition, there is "a kind of temporary variable binding" (pg. 9) that allows us to associate information with a particular spatial location. Another example Clark uses, is that of the associating of a particular motor routine with a visual cue, such as the way in which we look at an object as we go to reach for it. For Clark, this means that the external world is analogous to computer memory. In this way, interaction with the external world allows a thinking thing to be far more efficient in terms of internal memory requirements. In addition, the biological capability of an animal to make a feedback prediction before sense data actually arrives has been shown to aid in the refinement of motor capabilities and allow for mental rehearsal that is not linked directly to action. Mental rehearsal itself is a tool that allows for the improvement of a range of motor and cognitive skills, such as language and reasoning. Further still, interaction with our environment gives us access to physical tools and other agents, both of which aid in problem-solving.

In addition, in his 2024 paper "The Embodied Intelligent Elephant in the Room", Saty Raghavachary [Rag24] emphasises the need for models to directly experience the world through the physical body in order to generate human-level grounded meaning. It is only possible for models to associate symbols with direct interactions with the world via embodiment, coupled with a suitable architecture (he calls this "embrainment"). Models without this can only operate at a symbolic level and cannot experientially understand the physical world.

These points suggest that the ability to interact with and sense the physical world is a potential solution for the data efficiency crisis in machine learning. Fewer internal cognitive resources are required when the model has the ability to link more detailed information with a particular set of sense data, meaning that these detailed memory stores and calculations need not exist entirely internally. Moreover, these capabilities aid in the development of both physical and cognitive skills. This points to the importance for embodiment in aiding with data efficiency in model learning.

In addition to these ideas, simply having a continuous, multimodal stream of data and the ability to adjust one's perspective within the world helps to build understanding of that world a great deal more rapidly than, say, a DNN which only receives a series of static images that it must learn to categorise [Car+24]. Babies, for example, are able to reach out and grab, feel, bite, and throw objects to gain a deeper understanding not only of the properties of said object, but of the rules of physics governing that object. Intentionality comes into this when, as infants develop, they are able to choose the way in which they orient their sense organs and interact with the objects around them in order to fill specific gaps in their understanding [SH23]. For example, I am able to direct myself to look more closely at an object when I know that I do not yet fully understand its properties. This means that embodiment in addition to intentionality can further aid model learning. Further still, when models

can generate expectations of what they intend to learn, the senses can be more efficiently directed towards relevant sources of information. This combination of capabilities makes up embodied agency.

Receiving (sufficiently complex) multimodal data in this way, binding information to the external world, and being able to intentionally make investigations about the world would make it considerably easier for AI models to build a picture of its external world and make predictions about the outcome of its actions. We can conclude here that, given these capabilities, a model would need a great deal less data to develop cognitive skills such as adaptive and context-aware reasoning. As an additional note, the ability to interact with and learn from other agents is also a way in which data efficiency is improved.

1.3.2 Causal Reasoning

Causal reasoning also seems to be a natural consequence of agency, as in order to develop expectations, agents must have some sense of which sensations are caused by which particular actions. Being able to link actions to sensations gives at least an understanding of causation when the instigating event is an action carried out by the agent. This can come about because agents observe the consequences of a particular action and, when that same action is repeated, have an expectation of the same or similar outcome (this may require multiple instances for the connection to be made). As the agent experiments with a wider range of actions, they can develop a better understanding of causal relationships. When expectations do not match outcomes, we are able to update our expectation-generating assumptions to be more accurate.

With a sufficiently complex understanding of the world, an agent would then be able to observe similar processes as carried out by external agents or objects and make predictions about the outcomes of those events.

If an agent experiences uncertainty about the outcome of an action, they may also be able to intentionally test hypotheses. For example, if I had a desire to move a remote-controlled car but did not understand the controls, I can find out the outcome of pressing a particular button by repeatedly pressing it and observing the consequences. I can test other buttons to see if the result is different, and observe other agents pressing the button.

The ability to make predictions and test hypotheses allows me to go beyond observing that two events frequently occur one after the other and make conclusions about causation. The more observations and experiments that can be carried out, the better the agent will become at making predictions about new scenarios. These abilities allow models to climb Judea Pearl's hierarchy of causal understanding, moving from mere association to predicting the effects of actions, and finally to reasoning about alternate realities.

1.3.3 Transfer Learning

As previously discussed, agency may help to make transfer learning easier for AI models. However, this may only be the result of having more data about an environment, rather than requiring specific intentionality. Returning to the *Breakout* example, if a model has only ever played one version of the game before the paddle is moved, moving the paddle represents a paradigmatic shift in the world of the model. For humans, this is not the case. Having developed a general understanding of the laws of physics and the way these kinds of games operate, it is considerably easier for a human to adapt to the

new version of this game. If the model had perhaps played a wider range of games before or had more ways in which it was able to interact with the game environment, understanding the parameters of the new game could be made specifically easier.

This development does not, therefore, specifically require agency, but it does suggest an expansion of sensing and interaction capabilities. However, expectation generation allows for experimenting within the new scenario to make better predictions.

1.3.4 Self Awareness and Metacognition

As suggested by Dobbyn and Stuart (2003) [DS03], agency may lead to some amount of self-awareness. As discussed, agency (and some causal understanding) allows models to distinguish between which sensations are generated by their own actions and which are not. Further still, understanding which objects you have direct control over allows one to understand the boundaries of their own body and that they are a part of the external world that they inhabit. In addition, the ability to control your own perspective allows a being to place itself as an object within the world they inhabit. Being able to self-monitor and establish one's physical boundaries and capabilities allows for, at minimum, an understanding of what is and what is not part of and caused by one's self.

At high levels of agency, a metacognitive layer can also develop. By carrying out and observing the consequences of their actions, an agent may develop and internal understanding of their own physical and mental limitations. Understanding that an outcome was different from what was expected opens up the possibility for an agent to be able to reflect not only on what could have been done differently in terms of physical action, but also reflect on whether any reasoning or decision-making mistakes were made. With a sufficiently high level of awareness of their own cognitive processes, the ability to make mistakes makes it possible for an agent to reevaluate and improve those processes.

1.3.5 Caring

Perhaps the most hypothetical of the consequences we will discuss, agents that understand themselves to be agents existing in an external world may, if they develop a goal-directedness towards self-preservation, to some extent, develop a form of caring about the world around them. As beings in the world we care about what happens because we want events to align with our goals, be it survival, pleasure, avoiding pain, etc. I care about climate change, I care about climate change because I don't want the world to be destroyed. I don't want the world to be destroyed because it provides me with things that bring me pleasure and bring me closer to my desires. Furthermore, understanding other agents having similar goals allows for collaboration and empathy between agents.

Exactly what gives rise to emotions and whether or not an AI model could truly "feel" them, or "care" in the way that we perceive ourselves to care about things is an ongoing debate and one that would require multiple theses to discuss. For now, I simply make the claim that the self-awareness that may arise from a sufficiently high level of agency in a sufficiently complex external environment may give rise to model behaviours that emulate a desire for self-preservations and caring about what happens in the world.

1.4 Conclusions

In this chapter, I have outlined the types of cognition that have been described by philosophers and psychologists to identify the cognitive gaps that exist in current AI models. I have then and developed a view of agency as a spectrum which depends on how good the agent is at action selection and expectation generation, which is influenced by the sensing and interacting capabilities of the agent. Finally, I have explored how a high level of embodied agency can improve the data efficiency and reasoning capabilities of AI models, as interacting with the world has been shown to expand memory, having a multimodal stream of sense data has been shown to increase learning speed, and intentionality allows agents to direct their interactions towards the filling of particular knowledge and ability gaps. Furthermore, the development of agency inherently forms a rudimentary level of causal understanding which, given sufficient data and experience, could develop into more advanced causal reasoning skills. Transfer learning may also be made easier by expanding the investigative capabilities of a being. Following the writings of Dobbyn and Stuart, we also find that agency paves the way for a being to develop a basic sense of self-awareness, which may become stronger as agency also becomes stronger. A high level of agency may allow agents to develop metacognitive skills, as they can identify themselves and understand the thought processes that led to certain actions, which can then be evaluated. Finally, I have explored whether, by being an embodied agent existing with human agents in the world, an artificial agent might come to care, or at least seem to care, about the world around them and prioritise self-preservation. In the last chapter, I will explore agency and embodiment in AI models from a technological perspective.

Chapter 2

Having established that embodied agency has the potential to aid in overcoming a great deal of the limitations faced by modern AI models, it is now time to dive deeper into the relationship between embodiment and agency. This chapter aims to make the claim that a thinking thing must at the very least believe that it is embodied in order to have agency but that embodiment specifically gives thinking things agency over the physical world. I will use the term 'thinking thing' as a broad placeholder for a being that thinks because I aim to make as few stipulations as possible as regards the cognitive capabilities of thinking things as we discuss what it means for a thing that thinks to be embodied. We will begin this chapter by exploring and refining a definition of embodiment, as informed by modern literature on the topic. I will then do the same for the definition of agency. Once conditions for both characteristics have been determined, we will be able to example the compatibility of the two and explore edge cases to determine if there are scenarios in which they do not align.

2.1 Defining Embodiment

We now turn to understanding what it means for something to be 'embodied'. This meaning will be slightly different from 'having a body', which we will see throughout the course of this discussion. By the end of this section, I intend to have generated a short list of minimal requirements for considering an agent to be embodied.

2.1.1 What is a Body?

Rene Descartes, for example, distinguishes sharply between *res cogitans* (thinking thing) and *res extensa* (extended thing) [Des08]. Related to his famous "cogito, ergo sum", he defines the mind purely as a thing that thinks, without the requirement of a body, something that is defined by spatial extension and mechanistic properties. For Descartes, the mind, and therefore all cognition, is not physical at all, but it does control the body via, he believed, the pineal gland in order to interact with the physical world. This dualistic way of thinking about the mind and body is largely considered outdated by philosophers of cognition, as simply being an 'extended thing' is not enough for most philosophers to consider something to be a body. It is true that in order to have a body, one must exist as a physical entity, but as physicalism is now the most widely accepted theory in philosophy of mind, most philosophers would argue that having some physical extension is a requirement of being a thinking thing, as it is by physical processes that cognition takes place.

However, we do not consider an LLM like GPT-4 to be embodied or to have a body at all, although it does

possess physical connections constructed from cables, wires, physical memory storage, motherboards, etc. These physical objects are purely what allow GPT-4, and other machine learning models like it to carry out tasks. So, what, then, distinguishes this kind of physical extension from the kind of physical extension that leads humans to be able to describe our own bodies and engage with the physical world? Two answers come to mind here. The first is that my body allows me to experience the world. The second is that I have the ability to manipulate my body through movement.

2.1.2 The Sense Requirement

Let us begin by examining the first of these answers. In his *Phenomenology of Perception* [Mer12], Maurice Merleau-Ponty writes (pg. 92) "I observe external objects with my body, I handle them, inspect them, walk around them. But as for my own body, I do not observe it: to do so, I would need a second body, which would itself be unobservable." According to Merleau-Ponty, the body allows us to observe and investigate that which is not part of our own body, and that we cannot observe our own bodies in the same way. In short, to have a body is to have the ability to sense the external world in some ways. Our bodies are not simply containers for the mind. One might respond to this conclusion with a thought experiment similar to that described by Avicenna (or Ibn Sina) in his eleventh-century work 'On the Soul' [Avi06]: suppose that an adult human came into existence floating in space, with no ability to see or hear, and posed in such a way that no part of their body touched another part, and that they were not subject the feeling of any winds or supporting forces. This person receives zero sense data, but they would still be considered to have a body. This thought experiment is employed by Chalmers in his 2023 paper 'Does thought require sensory grounding?' [Cha23] as an example of a 'pure thinker'. However, in this thought experiment our 'pure thinker' still has sense organs, they have just found themselves in a situation in which they are not receiving sense data, so they still meet the criterion of being able to sense the external world in some way, even if they never actually do so. Instead let us consider an altered version of this thought experiment: suppose a person is born without any senses. This 'senseless thinker' is blind, deaf, has no sense of smell or taste, and they have no feeling in any of their body parts, this person not only does not receive sense data (like the 'pure thinker') but they also *cannot* receive sense data of any kind. In this scenario, while the thinking capabilities (i.e. the brain) of this person still reside within a container with only trivial visual differences from a human body, I would argue that such a person is not embodied, nor do they truly "have" a body. This is because it makes no difference to the experiences of this person whether or not they have sense organs or not. If the technology existed, this person could have their brain removed and kept alive in a lab and they would not know the difference. The physical extension of this thinker beyond the brain is inconsequential to them. It would be the same as attaching googly eyes to a server tower. Thus, we can conclude that a necessary condition for having a body is the ability to sense the external world. We will discuss later in this chapter how we should define what it is to sense.

The 'sense requirement', as I shall call it, for having a body runs somewhat contrary to the writings of many philosophers. For example, Edmund Husserl [Hus52] makes a distinction between the lived body (*Lieb*) and the physical body (*Körper*). *Lieb* is the source of our intentional experience, and *Körper* is the physical body that can be experienced as a thing in the world by another agent. In introducing the sense requirement, I have differentiated three cases: *being* a body, *having* a body, and being *embodied*. The first is the same as Husserl's *Körper*: a thing (at least according to physicalism) *is* a body as it is

an object with contiguous material existence in the physical world. A body need not think. *Having* a body, by contrast, is similar to *Lieb* in that the body is able to "experience worldly things" [Weh20]. This is where the sense requirement, as I have described it, comes in: the thinking thing must be able to sense in some way, as the physical components of the body must make some experiential difference to the thinking thing. The details of this distinction will be discussed further in the section on defining senses, as naturally the physical makeup of a machine learning model's circuitry makes a difference to its output capabilities, but for now we will simply refer back to the word "worldly": the difference between "being" and "having" a body is made by the ability to directly access information about the state of the external physical environment.

2.1.3 The Manipulation Requirement

Unlike Husserl, I argue that there is a third distinction to be made. To be *embodied*, as opposed to *having* a body, one must also be able to intentionally interact with the world. An embodied thing must be able to direct their senses towards and manipulate parts of the physical world. [1]

However, there are ways in which we might cause problems for this criterion. Let us propose another thought experiment. Consider a person with Locked-in Syndrome (LiS), a rare and serious neurological disorder in which part of the brain is damaged [Sch+23]. People with LiS have total paralysis, but retain their normal cognitive abilities and senses, which means they can still see, hear, taste, smell, think, and feel. Some patients with LiS also retain some proprioception and sensation throughout their bodies. In this thought experiment, we have an individual who cannot move at all, but retains some sense ability. By meeting the sense requirement from the previous section, we would say that this person *has* a body. But are they *embodied*? Most people with LiS are able to communicate via eye movements and assistive technology, so this meets the movement requirement. But suppose that this individual cannot move their body at all, including their eyes. Without the kind of mind-reading technology that does not yet exist (such possibilities will be discussed in Chapter 3), this person *has* a body but is not *embodied*. They are not able to interact with the external world in any way at all. Their body allows them to receive inputs, but not to intentionally produce any outputs. Even an individual existing in a vacuum such that they can never make contact with a non-body object can produce outputs because their body is a part of the physical world that they can control.

Embodiment, as defined for this discussion, requires the perception of inputs and the production of outputs.

2.1.4 Being a Body, Having a Body, and Being Embodied

Let us formalise what we have so far.

To be a body, a thinking thing must:

- Be a contiguous collection of matter that exists in space and time (The Physical Requirement).

To have a body, a thinking thing must:

- Be a body and

[1] We might be able to reduce this to simply a requirement of being able to intentionally cause bodily movement, whether or not this is the case will be discussed later in the chapter.

- Be able to sense the external world (The Sense Requirement).

To be embodied, a thinking thing must:

- Have a body and
- Be able to directly control the movement of some objects in the external world (The Manipulation Requirement).

2.1.5 Contiguity

The only criterion that we have not closely examined thus far is the Physical Requirement. Being physical seems like a trivial requirement, but need a body be contiguous? It would seem at first glance that it does, as this part of the physical requirement helps us to determine what is and is not part of one's own body. This requirement is important because it gives us a clear distinction between what is and is not the body of a particular thinking thing. It also prevents us from making the claim that a particular thinking thing has more than one body. It does not prevent the case in which two thinking things may occupy the same body, but this is a strength of the criterion, as it allows for cases such as conjoined twins.

However, we may raise the case that contiguousness with the thinking matter is not a necessary condition for that which may allow the thinking thing to fulfil the Sense and Manipulation requirements. Let us use movement (which allows for the fulfilment of the Manipulation requirement), as an example. Intuitively, some objects that are part of my body, such as my hands, can be moved directly simply by thinking about doing so (thinking about moving is not itself a necessary condition for movement, we will discuss this further in the section about the difference between reflex and action). In order to manipulate an object that is not part of my body, there must be an intermediary step in which I move a part of my body (making noise, activating motion sensors, throwing items, generating vibrations, etc).

While we do not have the direct ability to consciously move all parts of our bodies (I cannot, for example, directly compel my liver to move), such things are contiguous with the parts that I can move. For example, if I were to move my legs and travel from one room to another, my liver, and the rest of my internal organs for that matter, must come with me.[2]

What, then, if there exists an object that I can move directly through thought, but that is not contiguous with the matter of my brain? Suppose, for example, that I develop telekinetic abilities and am able to move objects on the other end of a room simply by thinking about them. To give a more plausible example, smart home devices have the ability to remotely control devices such as lights, fans, televisions, etc. without any movement from the device itself. If such a device had no other means of interacting with the external environment other than through the direct manipulation of non-body objects, would such a device be considered embodied? What about if those objects could be controlled by multiple thinking things? One might make the point that in the case of a device with the capability to remotely control other objects there is some kinetic chain linking the two. In response, I would refer back to the first thought experiment, in which the manipulation is the result of some unexplained telekinetic power. In either case, the thinking thing in question (assuming it also fulfils the Sense

[2] I would concede here that something like a splinter is temporarily a part of my body.

Requirement) is considered embodied under the definition as I have given it. It is my claim that being physically contiguous with those objects which permit sensing and interacting is not a requirement for embodiment, and thus that the conditions I have outlined are both necessary and sufficient for a thinking thing to be considered embodied. This particular claim will become important in the next section of this chapter, where we will investigate the definitions of action and agency.

This conclusion aligns with Clark and Chalmers' [CC98] *Extended Mind Hypothesis*. They argue that the separation between mind, body, and environment is entirely arbitrary. For them, if an external tool (such as a calculator or notebook) plays the same functional role as an internal cognitive process, then it should be considered part of the mind. Other philosophers, such as Walker and Sparrow [WS24], develop an analogous argument (*The Extended Body Hypothesis*), which claims that bodily processes are not confined within the skin but extend into the environment. This theory, while originally developed for the discussion of devices like respirators and mobility aids, supports my argument that the objects a thinking thing uses to sense or interact with the world need not be contiguous with the objects that perform the thinking. Sensing and manipulating are bodily processes (if only in that they are conditions of embodiment), and whichever objects that serve this purpose are part of the (extended) body of the thinking thing.

2.2 Action and Agency

Let us begin with a more detailed investigation of the concepts of action and agency. In the previous chapter, I defined agency as the capacity to act (interact with the external world) intentionally (with an expectation of desired outcome) and to compare the outcome of that action to the expected outcome. This is the definition that I will argue for in this section.

The generally accepted understanding of agency is that it is the "exercise or manifestation of this capacity [to act]" [Sch19], which aligns with my definition. Given this definition, it is pertinent to first clarify what we mean by 'action' and what it means for a being to 'act'. I will break down what intentionality is and make the case for why expectation is a necessary component of intentional action. After understanding this, we can proceed to determine how this relates to my definition of embodiment. In the final part of this section, I will define what it is to have what I will call 'a sense of agency' and explore some evolutionary arguments for how this could have developed in biological creatures.

2.2.1 Defining Action

The colloquial definition of an 'action' is simply that it is a thing done, but for our philosophical purposes we additionally require the stipulation of the existence of some kinds of mental states in relation to the thing done. Firstly, this is so that we can define an actor as something that has the capacity to perform an action. Without this stipulation, an actor could be any object that is part of a causal chain (for example, a leaf that moves in the wind). Additionally, this stipulation allows us to differentiate between an action and a reflex, as, for example, an involuntary movement would not be considered a meaningful interaction with the external world.

In his 1953 *Philosophical Investigations*, Ludwig Wittgenstein [Wit53](§621) asks the question "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?". Many have read this question as suggesting there is something more to an action than the physical outcome.

For example, there is something that is different about me raising my arm, than me sneezing. The straightforward answer to this is that there is some form of intentionality involved. I chose to raise my arm, whereas I did not choose to sneeze. Even if I chose to do something (inhaled some pepper, for example) to make myself sneeze, the sneeze itself was a reflex that I used an act to trigger. I cannot think to myself "I am going to sneeze now" and then cause myself to sneeze with no other external input.

A very widespread theory of intentional action is known as the Causal Theory of Action (CTA), pioneered by Donald Davidson [Dav63]. According to CTA, something counts as an intentional action in virtue of its causal connection to certain mental states [PT23]. For Davidson, a true action statement denotes an event, and more specifically, a bodily movement. Such a bodily movement is an action if, and only if, the event is an intentional action. For Davidson, an event is an intentional action if, and only if, there is a "true rationalising explanation". According to many proponents of CTA, intentional action is explained in terms of mental states, as they are the "causal antecedents" of an agent's behaviour or bodily movement.

Part of this theory helps to explain why we can describe the consequences of an intentional bodily movement as intentional actions themselves. To use Davidson's example, suppose I flip a switch, and in doing so I turn on the light and also alert the burglar. On Davidson's view "flipping the switch", "turning on the light", and "alerting the burglar" are all descriptions of the same action: they are the same bodily movement, and said bodily movement is caused by a mental state.

Exactly which mental states are relevant for intentional action is a topic of much debate. Intentions seem to be the obvious choice here, as beliefs and desires do not appear to directly control bodily movement in the same way. However, Davidson pointed out that simply describing actions as intentional if, and only if, they are caused by some set of mental states leads us to 'The Problem of Causal Deviance'. There are two cases here that Davidson [Dav73] discusses that he calls 'Primary' and 'Secondary' deviance.

Secondary Deviance is when CTA would misdescribe one of the consequences of an action as intentional when it was not, such as if an agent wishes to kill someone by shooting them. They miss, but the target dies anyway due to a stampede that was triggered by the sound of the shot. While the outcome was intended by the agent, the steps leading to the outcome were not. The agent intended for the target to die as a result of being shot, which is not what happened. Had the agent taken the shot with the intention of triggering the stampede and killing the target, then the death of the target would be considered intentional. This indicates that intentional action involves some kind of plan for how the outcome should be achieved.

Looking at Primary Deviance can strengthen this conclusion further. Primary deviance occurs when we have an intentional causal chain that leads to an event which is not intentional. Returning to Davidson's initial example, suppose I did not know that there was a burglar in my home, then while "flipping the switch" and "alerting the burglar" are considered to be different descriptions of the same action, I intended for the former to take place and not the latter. So, the issue for CTA as a whole is that the causal chain alone from an intentional mental state to an event is not sufficient for the event to count as an intentional action. The difference between "flipping the switch" and "alerting the burglar" appears to be that there was some amount of foresight involved in flipping the switch. In

lifting my arm to make contact with the switch, I had the expectation that the switch would move under the pressure applied, and that the lights would come on. Given my lack of information about my environment (I was not aware of the presence of the burglar), I was not aware that this action would lead to a burglar being alerted, so this event was not something that I expected to take place as a result of my movements.

Consider a different example. Suppose I am playing basketball with my friends. When I receive the ball and throw it at the basket, I expect (or, depending on my abilities, I hope) that the ball will travel through the air at a certain trajectory such that it falls into the basket. This given my intention, I have used my cognitive and physical abilities as best I can to move my body in such a way that makes it likely that the ball will travel as intended. If, instead, the ball bounces off the backboard and hits my friend in the face, this was an unintended outcome. Intention is not just about a particular mental state that causes an action, but about the expectation that I have for how events will transpire as a result of that action. The unintended outcome of missing the basket, or of alerting the burglar are a result of limitations in my physical and cognitive abilities, and/or of information that I have about the external environment. These are factors that limit my ability to generate accurate expectations of the outcomes of my actions. By contrast, an omniscient and omnipotent agent would never experience an action leading to an unintended outcome, as they would always be able to perfectly predict the consequences of their movements and precisely guide their actions towards the intended outcomes.

An important aside here is that for an act to be intentional it is not required that the agent be made aware of the entire causal chain of events. For example, if I threw the basketball blindfolded and never found out whether it went into the basket or not, the event of the ball going into the basket still an intentional action.

What I have highlighted via these examples is that intentional action is not just an event that the causal consequent of a particular mental state. The particular event must be part of the expectation that the agent has for the outcome of their action. If the same desired outcome is achieved, but via an unexpected causal chain, it is not intentional. If an unexpected additional event is caused by an action, it is not intentional. My expectation is the "best guess" I have for the outcome of my action, based on my understanding of the external world and the level of control I have over my bodily movements. My expectations aid me in selecting the best plan for fulfilling my desires.[3]

2.2.2 How are Expectations Set?

In the previous section, we concluded that an intentional act is comprised of an action, an intention (a mental state that causes the action), and an expectation (a set of sensations the agent expects to experience as a result of the action). Now that we have constructed this understanding, and therefore have determined working definition for agency (the capacity to act intentionally) and an agent (something with agency), we can dig deeper into what exactly an 'expectation' is and how they are set.

In his 2016 book *Other Minds*, Peter Godfrey-Smith [God16] discusses the concept of an "efference copy" and reframes it in an evolutionary context. When the brain sends out a signal to initiate

[3]There is still massive ongoing debate in the philosophy of action that I will not be able to go through in detail here. Instead, I have highlighted the theories and discussions that are relevant for my argument.

movement, a duplicate of this signal (called a "motor command") is simultaneously sent to the body's sensory systems, this is called an efference copy. This allows organisms to generate an expectation of the sensory outcome of their bodily movement faster than they can receive actual sense data, which aids them in making motor adjustments in real-time. For example, I know that by turning my head I will feel my hair move against my neck, so I expect to feel this and am not surprised by this sensory input and I do not assume that I have been accosted by some other creature.

Godfrey-Smith makes an evolutionary argument that efference copies allowed organisms to develop an understanding of the bodily self. He argues that early animals would have benefitted evolutionarily from the ability to "tag" and filter out any sensory feedback that was the result of self-initiated movement. An efference copy allows an organism to sort sense data into two categories: internal origin and external origin. Sense data of internal origin is the result of the agent's bodily movements or functions (for example, the proprioceptive feedback of moving). Sense data of external origin is caused by something external to the agent. The result of such an ability is that an organism would be able to ignore predictable consequences of movement and focus on new and important sensory information, such as the rustle of leaves, which might suggest the presence of a nearby predator. Sense data that is expected can be ignored, and sense data that is unexpected is used to inform the organism's next actions.

The idea of the presence of a feedback loop that allows models to learn from and adjust their actions by comparing the sense data they expect to receive to the sense data they actually receive is supported by the work of Wolpert, Ghahramani, and Jordan [WGJ95]. By asking participants to estimate the position of their hands after moving in the dark, they found evidence that the central nervous system uses internal models to simulate the motor system. Forward models predict the sensory consequences of motor commands, and inverse models estimate the motor commands needed to achieve a desired outcome. Through repeated movement and sensory feedback, an organism can come to reliably predict the sensory feedback that will result from a particular motor command. If the sensory feedback is different to what is predicted, an organism benefits from being able to detect this mismatch and becomes better at determining which motor command to employ to achieve a particular outcome.

Godfrey-Smith additionally argues that, over time, as nervous systems became more complex, the phenomenon of efference copies would have contributed to the emergence of more sophisticated concepts such as selfhood and agency. This jump from basic efference copies to the kind of detailed outcome predictions achieved by humans is outlined by Rick Grush's [Gru04] emulation theory. He argues that the brain constructs neural circuits that model the body and environment. These "emulators" use efference copies to generate expectations of sense data. These models can also run offline in order to evaluate the consequences of different actions. With them, an agent can mentally simulate motor processes and perception because they can form expectations of, and interpret, sensory input.

As the complexity of these emulators increases, it stands to reason that some organisms would develop the ability to make increasingly detailed predictions for longer causal chains. For example, in addition to predicting that when I turn my head I will feel my hair move against my neck, I am also able to predict that I will be able to see further down the street and, because I just heard her call out, I will see my friend walking towards me. I may also predict that if I smile and wave at her, she will smile and wave back to me. I am able to make predictions using my internal model of physical laws and of social

behaviours. I am even able to make predictions about what would occur even when there is no chance of receiving particular sensory feedback by imagining what I would experience as an observer, such as in the blindfolded basketball example, where I would not be able to see the outcome of my actions.

The neurological phenomenon of efference copies laid the evolutionary groundwork for the human ability to make complex predictions about possible future events and their causal outcomes. This ability allows us to learn and make adjustments when our expectations do not match the reality. It is by lining up a desired outcome with a motor signal that we predict is most likely to cause this outcome that we are able to act intentionally.

Further still, Giovanni Pezzulo [PC09] [Pez11] argues that theses internal models and mental simulations give rise to particular kinds of knowledge and reasoning. He argues that procedural knowledge (knowledge of how to do something) is supported by the sensorimotor feedback loop. Declarative knowledge (knowledge of facts) arises from offline simulations, which allow for more abstract reasoning and planning. What began as a system of planning and learning from actions can also be used for cognitive functions such as memory, language, and imagination. This argument, if correct, demonstrates that not only is expectation generation essential for agency, but it is a vital tool that assists in the development of symbolic representations and complex reasoning skills.

2.2.3 van Hateren's View

In his 2014 paper 'The Origin of Agency, Consciousness, and Free Will', J. H. van Hateren [Hat14] makes a similar evolutionary argument for the origins of agency that also relies on the existence of a feedback loop. He begins by introducing two functions: f_{true} (fitness) and f_{est} (self-estimated fitness). These are predictive functions of time that track the potential success of an organism. Van Hateren defines success in this paper, in its most basic form, as the expected number of offspring of an organism, but notes that for a more complex social and cultural species, success would also include indirect components such as the success for related organisms. When the environment changes for an organism, it can respond accordingly via behavioural change. Not changing behaviour could result in death and is therefore a poor strategy. However, f_{true} is not directly available to the animal as it is not observable and could only, perhaps, be calculated by means of simulating the animal and others like it in its environment. This is where self-estimated fitness, f_{est}, comes in. This estimate could be based on a wide range of internal and external signals that might be available to the animal. By evaluating the outcome of a particular behaviour in reference to f_{est}, an animal is able to determine suitable behaviours given particular environmental factors, the outcomes of which can then themselves be evaluated. When fest is low, there is more variation in behaviour than when it is high.

Van Hateren argues that this loop of behaviour determination and outcome evaluation is the origin of an elementary form of agency. The variance in behaviours is determined by f_{est}, meaning that the behaviour of the animal is at least partly determined by sensory feedback. Additionally, there is an intrinsic goal-directedness in the animal, as f_{est} is part of a control loop that implements agency, it is a goal of the organism to have a high f_{est}.

Through the evolutionary accounts of both Godfrey-Smith and van Hateren, we see possible explanations of the development of agency in cognitive organisms. In both cases, agency develops via a feedback loop based on received sensory information and continuous adjustment over time. Agency comes in

when organisms develop the ability to use their generated expectations of sensory feedback to decide which course of action most aligns with their goals and desires. These descriptions align with our earlier definition of an intentional action, in that behaviours are enacted towards a goal and there is an expectation of sensory feedback that is compared with a real outcome. This feedback is then used to inform and refine the decision-making process for future actions. Therefore, expectation is both definitionally important in order to properly determine what is an is not an intentional action, and it is itself a cognitive skill that can be refined and improved upon through experience.

Moreover, these accounts point to my previous assertion that agency as a spectrum of capability rather than something that an animal does or does not possess. The level of agency that an agent possesses is determined largely by their ability to receive sense data, the accuracy with which they can make predictions about the consequents of bodily movements, and how well they can carry out such movements. Agents with a high level of agency are able to receive and interpret a great deal of sense data and use it to make accurate and detailed predictions. These agents can then use those predictions to select actions that align with their goals with a high level of accuracy. We might say that high-level of agent has a high level of control over their external environment, as the understand it in great detail and can use this understanding to plan courses of action that alter the state of the environment to precisely match their desires.

2.3 Embodiment as a Condition of Agency

Now that we have taken the time to outline and argue for our definitions of both embodiment and agency, we can begin to answer the question of whether embodiment is a condition for a thinking thing to have agency. I will argue that an agent must at least have a sense that it is embodied.

At first blush, the connection between the two seems almost definitional. Based on the discussion so far, an agent has the capacity to act intentionally. As we have defined thus far, an intentional act is a bodily movement that is the expected consequent of an intentional mental state. In order to make a bodily movement, an agent must be a body, and therefore they must meet the physical requirement. If an agent can move intentionally, then they meet the manipulation requirement. Finally, if an agent has expectations of sense data, then this implies that they are capable of receiving sense data, and therefore they meet the sense requirement.

I note here that I have not included mental actions in my definition, such as the 'action' of adding 2 and 4. I do not claim that mental actions are not actions, but I do claim that in order to be a full agent, a thinking thing must be capable of an *external* action (an action that manipulates the external world in some way). As I will argue later, I make this claim because a thinking thing that cannot act externally cannot compare the expected consequences of their actions to reality. If the act exists entirely within their mind, so do the consequences, meaning that the feedback loop collapses in on itself. There is nothing to distinguish the generation of an expectation from the act itself[4]. Thinking through what the result of a particular sum might be is the same as performing the calculation, and then there is no way of checking this answer. I leave a discussion of agency with regard to purely internal actions outside the scope of this thesis.

[4]Philosophers such as Alvin Goldman [Gol06] argue that simulation is a central cognitive mechanism, further reinforcing the idea that expectation generation is itself a mental act.

However, let us examine these conditions more closely and consider whether all three embodiment conditions are truly necessary for a sense agency. As far as we know, all objects that think are physical, so the physical requirement is already met. However, for an agent to have a sense of embodiment, they need not actually be physical. So, this requirement is not important, even though it does appear to be contingently met.

2.3.1 Are all Acts Physical?

We begin by reviewing the manipulation condition. The manipulation condition states that in order for a thinking thing to be embodied, it must be able to intentionally control objects directly. As the body is physical, the implication is that control means the ability to move parts of the body. Thus far in the discussion, I have argued that an action must also involve bodily movement. If we can call into question whether this is the case, then we can question whether embodiment is truly a necessary condition of agency.

In his paper 'Agency and Embodiment: Groups, Human-Machine Interactions, and Virtual Realities', Johannes Himmelreich [Him16] describes three kinds of 'disembodied actions': Proxy Actions, Extended Actions, and Extended Movements. Proxy actions are acts that were not performed by the agent responsible. For example, a spokesperson giving a statement on behalf of the president is performing a proxy action.

Proxy actions still align with our working definition of an intentional act, as there must still be some causal chain of events from a bodily movement of the agent to the performance of the proxy action. In giving consent for a spokesperson to make a statement, the agent has made an intentional bodily movement and therefore the action is still embodied.

Extended actions are similar to those considered at the end of the first section of this chapter. Extended actions take place when an agent performs an action, but the agent's body does not move in performing that action. The example that Himmelreich uses is of an agent (whom he names Jim) who is able to control the movement of an object using some form of brain-machine interface, allowing the brain to directly send signals to a moving machine. As argued above, the objects that an agent has direct control over need not be part of their contiguous body in order for the agent to meet the manipulation requirement. David Armstrong [Arm68] even goes as far to say that the body is just those objects under an agent's control[5], regardless of how they are connected to them. Therefore, even an agent that can only perform extended actions (say, a person with LiS connected to a brain-machine interface) would meet the manipulation requirement, and therefore still be considered embodied.

Finally, we have Extended movements. Extended movements occur when "[the agent]'s body moves in performing [the action] but [the agent]'s narrowly individuated body does not move." [Him16] (pg 10). These movements can also be explained as embodied if we do away with a contiguity requirement, as we have done with the case of extended actions. The example of Jim does not change at all. In fact, Himmelreich does not make a clear differentiation between extended movements and extended actions at all if, as he suggests we might, we give up the idea that the objects over which an agent has

[5]This would imply that something like a hammer is part of the body as long as it is in use, which aligns with the claims of the *Extended Body Hypothesis*. Examples like this are not particularly relevant for this discussion as a tool like a hammer still requires physical movement to be used.

control must be part of their contiguous body. By doing so, I have done away with the possibility of extended actions entirely and defined extended movements as embodied. However, the discussion of extended movements does give rise to one final edge case that may yet cause issues for the manipulation requirement.

Suppose that a patient with LiS is given access to an artificial humanoid body which they can control directly. Then the patient is using this body (or avatar), they perceive everything as if they were the avatar, with no conscious awareness of their biological body. This scenario is still no problem for our definition of embodiment, for reasons we have discussed. The objects that the agent uses to receive sense data and the objects the agent can control directly need not be contiguous with their body.

But what if this avatar were purely digital? Consider the famous 'brain in a vat' thought experiment [Put81]. If a thinking thing were only able to perceive digital data, believing itself to be an embodied creature in an entirely simulated universe, would we call this creature embodied? My answer is that we would not. The manipulation requirement stipulates that a thinking thing must be able to sense and directly control objects in the physical world. In this case, the thinking in thing is not embodied, it just believes itself to be.

We might, however, still consider this being to possess agency. In terms of mental states, and agent performing an intentional action in the physical world is identical to a twin brain performing the same action in a twin simulated world. We might save the manipulation requirement by stipulating that an agent must somehow be deemed so by some other agent, but this moves us far away from understanding agency as a cognitive capability and towards something altogether different. Such a discussion is beyond the scope of this paper.

What, then, differentiates my intentional actions in the physical world from those in, say, a very vivid dream? If the two are the same, then this would lead us to conclude that thinking about throwing a ball is the same as actually throwing the ball. If this is true, then any link between embodiment and agency entirely collapses. The difference, I argue, is that the external world is just that: external. Even in the simulation scenario, the simulated world is not generated by the brain itself. Those signals that lead the brain to believe that it is embodied and performing actions in the physical world are externally generated. The agent still experiences themselves to be manipulating objects in the external world, whether or not that external world is actually physical. As with the physical world, the sense data an agent receives from an external world world can be generated by objects over which they have no control. By contrast, in a dream world, all parts of the world are generated by the agent's brain and are not external, meaning their actions are not external, they are mental. As mentioned at the beginning of this section, an agent must be capable of external actions, because there is no way for the agent in the dream world to be wrong about the consequents of their "actions". The generation of an expectation is identical with the act itself. I will argue this point in more depth in the next part. Therefore, even though the agent does not meet the requirements for embodiment, they do still have a sense of embodiment, in that they must still be able to interact with an external world.

This argument has made clear the difference between embodiment and a sense of embodiment, and why only the latter is required for agency. An agent at the needs to be able to interact with an external world, even if that world is not physical. Embodiment, however, stipulates that such interactions be with physical objects in the physical world. This reinforces the conclusion that a sense of embodiment

is required for agency, not embodiment itself. Being an embodied agent is what gives agents control over the physical world that humans inhabit.

2.3.2 Reviewing the Sense Condition

Does a being need to be able to sense to act intentionally? So far, we have had no particular reason to think that it is. Despite this, it seems unusual to imagine performing an action without any kind of sense feedback. For example, a person with paralysis can send the same motor signals to a paralysed limb that they did prior to the paralysis, but this does not mean that said signals will be received. Intention is not enough for agency. Even if a person were to suddenly lose all ability to receive sense data, thus still being able to mentally simulate actions and generate outcome expectations based on their experiences before this event, they would have no way of knowing if they performed the intended action.

We might make a distinction between internal and external sensations. Even the being that exists in a vacuum would experience some proprioceptive feedback, such as the feeling of moving their limbs, and would know that they had done so, allowing for the kind of feedback loop described in our evolutionary examples. Any objects that a thinking thing is able to control are themselves part of the external world, even if they are also part of the body of that thing. Therefore, internal sensations are also sufficient to meet the sense requirement for embodiment.

But what if the thinking thing were unable to receive any kind of feedback from an action, proprioceptive or otherwise (like the 'senseless thinker')? I would say that such a being would not be an agent. I make this claim for the same reason that I argued that agents must be capable of external actions: the feedback loop collapses without the ability to compare expected outcomes with actual outcomes. Agency develops through the repeated testing and adjustment of action plans, a process that sense feedback, whether internal or external is essential to. If an agent is unable to receive any kind of information about the consequences of their actions, they are unable to develop any kind of control over their actions or the external world. One motor signal is entirely indistinguishable from another (if these could even be generated without proprioceptive feedback), meaning that the agent cannot select actions for the pursuit of their goals because they have no way of knowing the possible consequences of their actions. A being that has never experienced any kind of sense data would not have the ability to form expectations (or intentions, for that matter) of any kind.

The case of an agent that has suddenly stopped receiving sense feedback of any kind is a little more difficult. They may have developed desires and the ability to act intentionally, but now have no way of knowing if and how they are acting, as well as whether the external world still behaves in the way they might expect it to. They receive no sensory feedback of any kind, including proprioceptive feedback. In short, this once-agent has lost any control that they had over themselves and the external world and they can no longer develop or refine their expectation-generating abilities. Their mental life becomes indistinguishable from a thinking thing with no ability to manipulate the external world – as far as the agent is concerned, it is only capable of internal action.

The reason I argue that an agent must have a sense of embodiment, rather than actually be embodied, is that regardless of whether it is simulated or comes from the physical world, sense data would appear identically in terms of the mental states of the receiver. The capacity of agency is therefore

unchanged. If the external environment is simulated, however, it must be simulated externally to the agent, otherwise any actions would be purely internal.

In the final chapter of this thesis, we will discuss the ways in which artificial intelligence technology might utilise internal monitoring systems, and whether a lack of external monitoring systems has significant cognitive consequences.

2.4 Conclusions

This chapter has aimed to explore the question of whether embodiment is a requirement for agency. In the first section of this chapter, I went into detail as to what exactly we mean by embodiment, and I argued that there is a difference between being a body, having a body, and being embodied. I argued that, in order to be embodied, a thinking thing must be able to sense the external world in some way and directly control objects within it. I made the case that the objects that a thing uses to sense and the objects it has direct control over need not be contiguous with the thinking parts of its body. In the second section, I argued that actions not only require intentional mental states but also require some form of expectation for the outcome of an action. I examined two evolutionary explanations for the development of agency in biological creatures. Finally, I examined whether embodiment is truly a necessary condition for agency. After exploring edge cases of the embodiment conditions, we reached the conclusion that while embodiment can often lead to the development of agency, there is a key difference between the two. Agency is a primarily mental capability, whereas embodiment is not. An agent must be able to interact with the external world, but for an embodied creature, that world must be physical. By contrast, both embodied creatures and agents must meet the sense requirement, as an agent must be able to determine that an action has taken place. From this we conclude that a being's *belief* that it is an embodied creature is a necessary condition for agency, but it need not actually be embodied.

Chapter 3

Having established that agency requires a sense of embodiment as well as, in Chapter 1, establishing that embodied agency has the potential to enable AI models overcome some of their limitations, we now need to understand what technological changes could be made to these models in order to make them embodied agents. We will begin with examining current research and ideas about how to make these models embodied, before turning to the more complex issue of intentionality (or something resembling intentionality) in these models.

There is already a great deal of research in the field of robotics [Rob24] [Cou24], but much of this research focusses on optimisation of the robotic mechanisms themselves, such as improving the sensitivity of cameras or refining movement capabilities. I wish to look at this problem from the perspective of how the capabilities of a complex AI model can be enhanced by the addition of *physical* tools for sensing and interacting with the external world. I will also discuss how the architecture or code of these models may need to be adapted to enable the use of these physical components, as well as the development of something that we might call intentionality. I will then discuss a range of model architectures that could be implemented for agency. I aim to demonstrate that there are agentive model architectures that, when combined with physical components for sensing and manipulating the external world, would result in a model that could be deemed an embodied agent. I will argue throughout this chapter that these physical tools and agentive architectures have the potential to produce models that can overcome some of the limitations outlined in Chapter 1. Focussing on specific individual additions that could be made to current AI models will help me to be precise about how exactly embodied agency has potential benefits for models.

3.1 Adding Senses

When we return to our three requirements for embodiment (the physical requirement, the sense requirement, and the manipulation requirement), we see that, since an AI model would already meet the physical requirement, there are two steps that need to be made for a model to be considered embodied: it must be able to sense the external world and interact with it.

We have already discussed in Chapter 1 that an LLM does interact with an external world (a world of user-generated textual inputs and probability-based outputs), but the mono-modality and limited input data of this world is part of the reason that these LLMs are limited in their capabilities. If we want LLMs to be capable of solving problems based in the physical world, such as causal reasoning or classification of real-world objects, then they must be able to sense and interact with it. For example,

work by Baravesco, de Heer Kloots, Pezzelle, and Fernández [Bav+25], found vision language models (VLMs), which learn from both images and text-skilled representations, that aligned more closely with human brain activity related to language processing than language-only models. This suggests that greater access to data about the physical world (in this case, multiple modalities) helps models to more closely mimic the brain processes of humans.

Of course, they could theoretically achieve the same capabilities by interacting with a simulated world that is just as complex as our physical world, but then the problem moves to how such a simulated world would be possible, as it requires a vast amount of computing power [Vid14]. So, for the purposes of our discussion, we will focus on the technological adaptations that can be made in order for AI models to *sense* and *interact* with our physical world.

The aim of this section is to explore ways in which models might come to meet the sensing requirement for the physical world and to demonstrate how these methods can help improve the performance or capability of existing models. Of course, models that only meet the sense requirement would not be embodied agents, but models that are embodied agents must meet the sense requirement.

This discussion will be split into three segments: sensing the external world, adjusting the senses towards particular objects, and manipulating external objects. With all three, a model would be considered embodied. I have chosen to make a distinction between the latter two segments, as some models may be able to control their own machinery (such as moving a camera), but not manipulate external objects. We will investigate the model architectures that may be required in both cases. In the next section, I will explore agentive model architectures.

3.1.1 Continuous Data

The most obvious place to start when it comes to giving AI models something akin to 'sense organs' would be with cameras, giving AI models access to a continuous video stream of information. When it comes to the task of image recognition, multiple similar frames help systems to recognise which parts of an image are redundant and also understand what an object may look like from multiple angles [Car+24].

It may be argued that these benefits are simply the result of providing a model with more data. However, what is important about a stream of visual data is that it is different from a series of discrete frames because it is continuous and therefore there is a relationship between those frames. A 2024 study by Shrivastava and Shrivastava [SS24] revealed that when video is treated as a "continuous multidimensional process rather than a series of discreet frames" (pg. 1), there is a 75% reduction in the sampling steps required to sample a new frame.

From a technological perspective, it is important to understand the changes that were made in order for the model to utilise the data effectively. In this particular study, the model was designed to make predictions about what the next frame of the video would look like. The framework created used a mathematical formula to create intermediate frames between video frames. Some noise was added to these intermediate frames to simulate natural motion and variability. The time variable was also modelled as continuous rather than discreet. As a result, the model was more efficient in its prediction.

This use of continuous visual data mimics the ways in which humans and animals sense the world –

instead of observing separate frames of a scene, we understand an event as a whole, fluid motion. By modelling time as a continuous variable, we unlock the potential for systems to recognise patterns of motion, observe how objects change over time, predict behaviour, and perhaps even begin to develop a rudimentary understanding of causality (in the sense of observing that certain events are always preceded by other certain events).

While this alone does not mean that a model is embodied or an agent, enhancing the sense capabilities of models that are agents (such as with continuous data streams) can lean to a higher level of agency.

3.1.2 Multimodal Data

Audio data, however, always has a time variable due to the nature of how it is received – any clip of sound has a duration. So, there are no obvious changes that might occur if a model trained on only audio data were to be solely equipped with microphones that were continuously receiving audio data, other than the expected benefits of simply receiving more data. So, while audio models may not benefit much from continuous data(any more than they already do with clips of sound), we might consider whether models can benefit from receiving multiple modalities of data.

When AI models receive both audio and visual data, they are better at making predictions. This was shown in a 2024 study by Su, Liu, and Shilzerman [SLS24], which found that models showed improved performance in speech recognition and object detection. The use of this kind of multimodal data allows models to understand the relationship between certain visual events and particular sounds. This means that models are better at identifying complex events, such as when there are overlapping sounds or multiple visual events. The *Contrastive Audio-Visual Masked AutoEncoder*CAV-MAE Sync model, developed by researchers at MIT and IBM, was able to predict sounds from video data and generate images from audio data [Gon+23] [Ara+25]. For example, it was able to match the sound of a roller coaster with the visual of its motion. This was shown to help with speech recognition in noisy environments and with emotion detection.

An understanding of the relationship between data from different sense modalities, which is created when that data is linked by a time variable, overall allows for a greater understanding of the real world environment. Not only are models better at understanding and predicting complex events, but they have a broader picture of the features of particular objects and scenarios.

3.2 Directing the Senses

Some of the senses we have not yet discussed include touch, taste, and smell. These differ from sight and hearing in that they more often require actions to be taken in order to be employed. In general, we choose to touch, smell, and taste a particular object by picking it up, inhaling through our nose, and putting it in our mouths. This is of course not always the cause. For example, one often finds that they cannot help but smell a particularly foul odour. The point is simply that our 'field of sensing', as it is, for sight and hearing is generally larger. Many objects that we can see and hear will need to be brought much closer, either by our own actions or by some external force, in order for us to also be able to touch, taste, and smell them.

Receiving important sense data about the external world is enhanced when we are able to move our

sense organs toward that which we wish to perceive.

The ability to direct our senses towards certain objects is part of what allows us to investigate the world around us. We move towards objects of interest and away from that which is deemed unimportant. We might call this skill 'attention', as there is some, however minimal (and perhaps unconscious), decision-making process which determines how the senses are directed.

3.2.1 Physical Directing

Before turning to the decision-making processes that may be involved in AI model attention, we can briefly mention the physical components that may assist with this. Firstly, some amount of camera movement and focus adjustment helps to adjust the visual field. A model might utilise this to follow an object of interest or scan a wider space for relevant information. In addition, the use of multiple microphones would help a model to identify where the source of a sound is in space by comparing the volume between each microphone. The camera could then be directed towards the source of the sound. Furthermore, the ability to manipulate objects would unlock even more information. If a model had access to a grasping mechanism to move objects around, the object could be observed from multiple angles, its movements examined, and its size, hardness, texture, and weight could potentially be determined. We will discuss further intentional manipulation of objects in a later section of this chapter. For now, we will move on to discussing how a model might determine where to 'focus' on its data receiving tools.

3.2.2 Attention Architectures

If a model is equipped with some device that allows it to be receiving raw, real-time data about the external world and has some goal-directed decision-making process that it uses to direct those devices towards particular stimuli, then we might be able to say that it has a rudimentary version of agency (the capacity to act intentionally with an expectation of outcome). It is able to sense the world and carry out actions (such as the movement of a camera) with an aim of gaining information relevant to its task. While it may not have fully simulated an expectation of the data it might receive in directing its sense devices, the model expects that it will gain the most useful information from the specific area it 'turns' toward. This means that, in simple terms, the model is turning away from the expected (that about which it does not need to learn more) and towards the unexpected. If a model is aiming to carry out a particular task, then it needs to be able to investigate new information. I would argue, however, that this is not necessarily enough for agency, as these behaviours may be simply reflexive and bypass any kind of symbolic reasoning.

In Chapter 6 of *From Deep Learning to Rational Minds* [Buc23], Buckner outlines a number of ways in which attention-like mechanisms can be utilised in ML model architectures. The first of these are called *Region-based Convolutional Neural Networks* (R-CNNs). These models, used, for example, for image classification, direct their resources towards areas of the image that are most likely to lead to identification, which leads to a significant improvement in efficiency and accuracy compared to traditional CNNs. The R-CNN uses a component called a *Region Proposal Network* (RPN) to compute a set of proposals for regions that should undergo more intensive feature detection. Within this type of model architecture, there is some (minimal) level of expectation in that the regions selected are expected to be most useful in the classification process.

The ability to direct the senses towards particular stimuli is another building block that could lead models to a high level of agency. This capability allows models to explore the external environment beyond the static receipt of sense data. Attentional models also form the foundation of the ability to interpret sense data. By selecting areas of importance, models are able to pick out the significant features of the objects around them. While this capability still does not constitute agency, an agent with sense-directing capabilities has a greater ability to receive and interpret data, which aids in the expectation generation process.

3.2.3 Recurrence and Continuous Learning

Buckner points out that R-CNNs are limited in that they lack recurrence, which is something that would bring a model closer to truly representing agency because being able to learn from sense feedback is very important for improving action selection. Recurrence in ML models is when past outputs are fed back into the model in order to inform the next output. For the R-CNN, the results of later processing do not influence the selection of the regions of interest. This, naturally, differs from human visual processing in that we will learn from one area of focus in a visual space and use this information to determine the next area of focus. For example, seeing that someone is looking in a particular direction might prompt me to also look in that direction.

Recurrence is a capability that is implied within agency, as agents are able to learn from the outcomes of their actions in order to develop better control over their actions. [1] Many models are trained on a specific dataset and then put to work on the specific task they were trained to carry out. A model with recurrence is able to alter its behaviour based on its outputs in order to make increasingly accurate predictions, even when the external environment changes.

Buckner then goes on to describe a *recurring attention model* (RAM), which uses information from a prior 'glimpse' of an input to determine where to focus next. This mimics human visual scanning more accurately, because the 'glimpse sensor' used is much more similar to the human retina, in that the resolution of the input increases toward the centre of the window.

In addition to mechanisms that adjust focus towards specific parts of a fixed visual field, there are also systems, sometimes known as *Active Vision Systems*, that use reinforcement learning or recurrence to control camera movement (such as panning or zooming) and maximise useful visual information. These kinds of system can be combined with audio input so that cameras can be moved towards target audio sources.

3.3 Manipulating the External World

We have now understood some of the physical and architectural systems that could be implemented to allow models to sense the world and direct those senses towards particular stimuli. We have also discussed some of the performance benefits that may result from this, in that multimodal and continuous data aid greatly in the efficiency and accuracy of categorisation tasks, and the ability to

[1] Theoretically, a model that acts intentionally with an expectation of outcome that never learns from the outcome of their actions would still be an agent, but just not a very good one, as they would be unable to adapt to the environment as it changes. The would therefore be at a lower level of the agency spectrum than models that do have recurrence, as they are worse at predicting the consequences of their actions.

focus on specific points of interest helps to identify and isolate particularly useful data.

We now turn to how an AI might be able to (intentionally) manipulate the world. This goes beyond the ability to adjust the senses, as it unlocks the potential to change the stimuli themselves and learn more about their behaviours. Although adjustment of one's own sensing organs / devices counts as an action, agents that are only able to act in this way are limited in the amount of information about the external world that they can access compared to agents that can manipulate the external world. As discussed in earlier chapters, agency exists on a spectrum and having a greater ability to manipulate the external world allows agents to achieve higher levels of agency. This is because the agent can gain a great deal more information about the external world via, e.g. hypothesis testing, leading to more accurate action selection and expectation generation, resulting in more control over the external environment.

In the coming section, we will discuss the physical components that a model might be able to employ to better manipulate the world. We will investigate possible model architectures that could be used to decide what physical actions to take, and finally we will talk about intentionality and how this might be at least simulated by an AI model.

3.3.1 Physical Mechanisms of Manipulation

There are a range of physical devices that can be controlled by AI models to manipulate the physical environment. The obvious examples that come to mind are the robotic equivalent of 'limbs', which give models the capability of reaching out and grasping objects in order to move them. There are also various ways in which the systems as a whole can move around a space, such as with wheels, bipedal, or quadrupedal locomotion. Much of this depends on the size and shape of the robot and the tasks it is expected to carry out. If the physical devices the controlling model has access to are designed specifically for a narrow range of tasks, then the model is likely to be less effective at transferring its skills to other tasks. Therefore, access to general tools that can be used to explore a range of objects and environments is beneficial for allowing models to investigate and try to solve a wide range of problems.

For example, Atlas (Fig. 3.1) is a humanoid robot developed by Boston Dynamics [Kui+15] with the ability to walk, run, and jump, as well as lift, carry, and organise objects. Atlas demonstrates the possibilities for models to be embodied with the physical hardware to perform a range of tasks and to have a high level of precision in their movements. These abilities further enhance the amount of information agents are able to uncover about the external environment. In addition, Atlas uses optimisation algorithms for planning and controlling its movements in real time, allowing it to perform complex movements in unpredictable environments. I will discuss model architectures like this in more detail in the next section, but this example at least serves as a proof of concept that recurrence-based planning models can control complex robots.

3.3.2 Proprioceptive Feedback and Reactive Architectures

Another element of sensing that is helpful in giving agents greater control over their actions is *proprioceptive* feedback, which is information about the movements and positioning of the body.

In a 2025 paper written by Zhang, Li, and Dai titled 'Continuous Learning and Adaptation of Neural

Figure 3.1: Boston Dynamics' Atlas Robot doing Parkour.[2]

Control for Proprioceptive Feedback Integration in a Quadruped Robot' [ZLD25], a four-legged robot was able to effectively navigate unpredictable terrain with the use of simple sensors in each limb that returned the angle of the Coxa-Femur joint (CF) of the limb. The CF joint angle refers to the angle between the coxa, which is attached directly to the body and usually controls horizontal rotation, and the femur, which is attached to the coxa and usually controls vertical movement. The control module in the robot is then able to compare the commanded and the actual CF angles and broadcast the "error" signal to all legs. Doing so allows the robot to anticipate any uneven terrain that is approaching and adjust the movement of the other legs accordingly. In this example, there is a feedback loop that closely resembles the way that biological organisms utilise efference copies [WGJ95]. A copy of the commanded motor signal is made and compared to the actual outcome of the movement. Typically, there will be a great deal more feedback that is available to a biological organism, but this kind of experiment shows that efference copies, a rudimentary form of outcome expectation, is not only a useful method for adapting to a changing environment but is also a very efficient way for robots to learn reliable patterns of movement, which can be used again in future scenarios. Further still, this kind of feedback can be combined with other sensory feedback, such as visual and auditory information, in order for models to make better predictions about the kind of motor signals that will be needed for particular tasks. For example, a model could learn to visually identify a particular type of terrain, such as a step, as it approaches and anticipate which motor signal will be needed.

Experiments [Li+25] with robots that receive proprioceptive feedback indicate that comparing expectation to outcome is an efficient way for embodied models to learn. Receiving real-time information about the status of the body aids embodied models in refining their physical abilities. As previously mentioned, the model architecture employed by Atlas displays this kind of real-time refining.

I note, however, that it remains to be seen if more complex agentive architectures can be integrated into these embodied robots. Some of the integration issues can be resolved by acknowledging that embodied agentive models do not need to be one cohesive "body", and could instead remotely control

[2]Image Source: Boston Dynamics

devices for sensing and manipulating, thus removing any restrictions on the size and shape of the thinking parts of the model.

3.4 Agentive Architectures

Having investigated how models might use immediate sensory feedback to adjust physical movements in real time (this is often called a *reactive architecture*), we can now turn to how this and similar processes can become increasingly complex for the pursuit of long-term goals and the solving of more difficult problems. Having the ability to solve problems that require more in-depth investigations requires a certain amount of pre-planning and understanding. For example, opening a puzzle box requires an understanding of what constitutes the box being 'open', the ability to guess what motor signals may lead to this result, and the ability to choose the next step based on the result of what was tried. This kind of reasoning is more detailed than simply comparing CF angles, meaning that a model needs to be able to develop a more complex understanding (and memory) of the external world. This kind of reasoning therefore requires more complex expectations and intentions.

This move from reactive to agentive architectures represents a shift from models that may simply be acting reflexively (and therefore are not agents), to models that use some amount of symbolic reasoning for action selection and expectation generation, which are agents. I will discuss these models in roughly chronological order, beginning with the earliest agentive architectures that relied solely on symbolic reasoning, and ending with some of the most recently developed models. At each stage, I will discuss the strengths and weaknesses of each type of model, arguing that more modern systems are our best options for high levels of (possibly embodied) agency.

There are a number of AI model architectures called 'Agentive AI Systems'. These types of system incorporate the elements that we have discussed so far in this chapter: a loop of perception, planning, action, and learning. But how do these kinds of model maintain and plan for these longer-term goals?

3.4.1 Deliberative Architectures

Deliberative Architecture Agents build and manipulate explicit world models and plan sequences of actions by filtering desires through feasibility checks and priority rules [GI89]. They then carry out these actions, monitor the results, and handle exceptions, as these models use explicit symbolic reasoning to calculate the best course of action.

One of the most influential approaches to automated planning is known as *Stanford Research Institute Problem Solver* (STRIPS) [FN71]. This is an early kind of deliberative architecture, and it is a planning system. States of the world are described by a set of logical propositions, and goals are defined as a set of conditions that describe the desired outcome. Actions in STRIPS are characterised as a set of three components: conditions that must be true for an action to be carried out, conditions that must become true as a result of the action, and conditions that must become false as a result of the actions.

Belief, Desire, Intentions (BDI) systems work slightly differently from this [RG91]. They maintain a set of beliefs which function as a knowledge base about the current state of the world that is updated based on sensory input. Desires represents the agent's goals and objectives. These objectives adapt based on the beliefs that the agent holds. Finally, the intention component represents the specific

courses of action that the agent plans to carry out. The beliefs and desires of the agent help them to determine which actions to commit to. Once the agent commits to an action, it will carry it out unless circumstances change.

Another type of deliberative architecture is a *Procedural Reasoning System* (PRS), which also has explicit representations of beliefs, desires, and intentions, but in addition it is equipped with a plan library, representing the agent's procedural knowledge, which is knowledge of the processes that the agent is able to employ in order to realised its intentions [GI89]. The PRS has an interpreter component which manages the beliefs, goals, plans, and intentions, continuously cycling between updating its beliefs and deciding what to do next.

Finally, there is the *Intelligent Resource Bounded Machine Architecture* (IRMA) [BIP88]. IRMA's structure is the most complex of the four deliberative architectures in this subsection. It has the same four symbolic data structures as PRS (plans, beliefs, desires, and intentions), but also employs some additional components. Sense data is fed into the belief structure, which communicates with a reasoner. The reasoner uses beliefs to make inferences about the world, and these are added to the belief set. Beliefs and plans feed into a means-end analyser, which determines which plans might be used to achieve the agent's intentions outputs these as options. Options are also generated by an opportunity analyser, which takes information about the environment and the agent's goals to determine which actions are possible. These options go through a filtering process to determine which options are compatible with the agent's current intentions. The surviving options, combined with the beliefs and desires finally go through a deliberation module to determine the best intention to adopt, and then the intentions are updated, being output as actions.

Deliberative architectures handle explicit planning well and adapt to novel situations by reasoning over symbolic models. They are particularly useful when there are strict restraints on the agent, such a safety regulations or correctness requirements. However, the use of explicit symbolic reasoning has a fair number of limitations. Firstly, it is computationally intensive. These types of models require a detailed internal model of the external world, which, for increasingly complex tasks, requires a great deal of computing power. Modelling the external world in this way also means that models can struggle in changing or uncertain environments, as the internal model may not always match with the actual world and there may be a delay in the updating of plans. Additionally, these models require highly accurate systems to convert raw data to symbols. As we have seen earlier in this chapter, if a model is receiving a constant stream of real-time, multimodal data, then this process not only takes a great deal of computing power, but requires a sophisticated system of object detection and identification, an advanced ability to make inferences based on this data, and being able to do so in enough time for those inferences to be useful (even if we employ attentional systems to improve efficiency). If we wish our models to be able to solve a range of complex problems in real time, then this type of system is likely not feasible, as the entire external environment is too complex to be accurately modelled symbolically.

The issue of computational complexity that arises here is related to what is known as *The Frame Problem*, which refers to the issue of making an alteration to a representation of the world, such as tracking the consequences of an action [MH69]. The problem was introduced in 1969 by John McCarthy and Patrick J. Hayes. In 1971, Hayes issued a memo proposing some solutions to the problem [Hay81].

One of the proposed solutions requires a more robust understanding of causal relationships, allowing systems to more efficiently infer the consequences of actions. In Chapter 1, I discussed the possibility for embodied agency to give rise to a higher level of causal understanding in AI models, so this is one possible way that this problem might be resolved. In addition, Hayes emphasises the need for efficient, context sensitive world models. Something like this may be achieved through the use of hybrid models, which will be discussed in the next subsection.

3.4.2 Reactive and Hybrid Architectures

In response to the difficulties with these symbolic systems, researchers have developed *Reactive Architectures* [Lee00]. These are behaviour-based models of activity, which are considered to be a lot more flexible that deliberative architectures because they can manage their resource abilities in unpredictable worlds. One of the earliest architectures that works in this way is known as a *Subsumption Architecture*, introduced by Rodney Brooks in 1986 [Bro86]. Brooks was a vocal critic of symbolic AI and argued that intelligent behaviour can be generated without explicit representations and explicit reasoning. The subsumption architecture that he proposed organises intelligence into parallel layers which run concurrently and can override each other. The lowest layer handles reflexive tasks and runs the fastest to ensure real-time responses. The middle layers build on these lower layers and are able to override the outputs from the lower layer if needed. The highest layers of this architecture perform goal-directed actions such as exploration and can override lower behaviours to achieve higher-level objectives.

While reactive architectures are great at real -time reactivity and flexibility, they cannot take past events into account or foresee the future. Their actions come only from perceptions, and thus they are difficult to predict and perform poorly at complex actions and pursuing longer-term goals. Therefore, some researchers have suggested that combining the capabilities of deliberative and reactive systems might help agents to react quickly to immediate changes in the environment, while still pursuing more complex objectives [Gat92].

Such a hybrid architectures are similar to what many philosophers and psychologists have theorised about human systems of decision-making, in that humans have two types of reasoning, one which is rapid, intuitive, and automatic, largely pertaining to making quick decisions in real time, and another which is slower and more analytical, pertaining more to the pursuit of longer term goals. This second type of reasoning can override out instinctive responses of the first type [Kah11].

One example of a hybrid architecture is a *Touring Machine* (not to be confused with a 'Turing Machine') [Fer92], which consists of a perception subsystem that feeds information to three layers: the reactive layer, which produces an immediate response; the planning layer, which manages typical running, assuming that the circumstances of the world are unchanged; and the modelling layer, which predicts conflicts and generates solutions to these problems. These layers are managed by a control subsystem, which decides which of the three layers should take control over the agent. The result of this is then fed to the action subsystem. This architecture is different from the subsumption architecture in that each of the layers may contain some explicit representations.

Another hybrid architecture is *Integration of Reactive Behaviour and Rational Planning* (InteRRaP) [MP94]. This model architecture uses a hierarchical knowledge base, which interacts with

hierarchical layers the control agent behaviour. The world interface, which consists of perception, communication, and acting components feeds into the world model, which is an internal representation of the external world. The world model interacts with the behaviour-based control component, which controls reactive behaviours. It generates and revises beliefs, activates reactions based on perception. It also generates requests to the plan-based component, the layer above, if needed. The plan-based component generates plans of action for the agent based on the plan library, which is the next layer in the knowledge base above the world model, and sends requests to the next layer up, the cooperation component. The cooperation component generates joint plans to satisfy the goals of a number of agents using cooperation knowledge, which is the top layer of the knowledge base and represents a social model of the world. Not only does InteRRaP unite reflexive and deliberative reasoning, but it is also able to do this with social reasoning about other agents by maintaining models of those agents, which goes beyond the capabilities of the models we have discussed so far.

While hybrid systems are able to manage both real-time problems and complex routines, they have their own limitations. There is the challenge of managing communication between layers and balancing resource allocation between them, because communication is multidirectional and must be timed and sequenced precisely. Further still, these models still face the problem of needing a great deal of computing power to internally represent the external world complexly enough to achieve challenging, long-term goals.

Finally, there is potential for a great deal of rigidity for having pre-designed plan libraries and patterns of behaviour – they cannot effectively update their strategies based on what they learn about the world [Woo09]. As argued in my discussion of recurrence, agents that cannot update their decision-making processes are likely to remain at the lower level of the agency spectrum, as they cannot refine their methods of choosing actions. It is true that hybrid architectures are an improvement in modelling agency compared to reactive architectures (actions are not simply reflexive) and deliberative architectures (they are better at reacting in real-time to changes in the environment, but the level of agency these models can achieve is limited because they cannot learn from the results of their actions.

3.4.3 Cognitive Architectures

We now move to an example of a more modern version of hybrid architectures: cognitive architectures. This is a very broad term for architectures that seek to emulate human cognitive processes such as memory, learning, perception, and problem solving.

Adaptive Control of Thought – Rational (ACT-R) (Fig.3.2, which is short for , is a hybrid cognitive architecture developed at Carnegie Mellon University [RTO18]. It differs from hybrid models such as InteRRaP in that instead of dividing functionality into separate reactive and planning tiers, it integrates all the cognitive processes with both symbolic and subsymbolic layers. In addition to declarative knowledge (facts about the world), which are stored symbolically, ACT-R has an additional memory module for procedural memory, which stores a series of IF-THEN rules that tell the model what actions to take if certain conditions are met.

ACT-R uses a set of buffers (goal buffer, retrieval buffer, sensory buffer, and motor buffer) to store the most recently perceived or retrieved information. When perception occurs, the sensory modules place new data in the correct buffer. The buffers are then scanned to determine if any if the procedural IF

conditions have been met. If so, the THEN action is activated. This might involve updating the buffers, retrieving information from memory, or issuing a motor command. The buffers are then scanned again and the cycle continues. If multiple IF-THEN rules match, then the expected value of each rule is calculated, and the rule with the highest expected utility is selected, although it does not have an explicit expectation of sensory outcomes.

What is ACT-R closer to high-level agency than the architectures in the previous subsection is that it is able to adapt its procedures based on sensory feedback. Each rule tracks how successful it has been and is less likely to fire if it has been shown to be less successful. In addition, complex sequences of rules are often combined into more efficient rules to speed up behaviours.

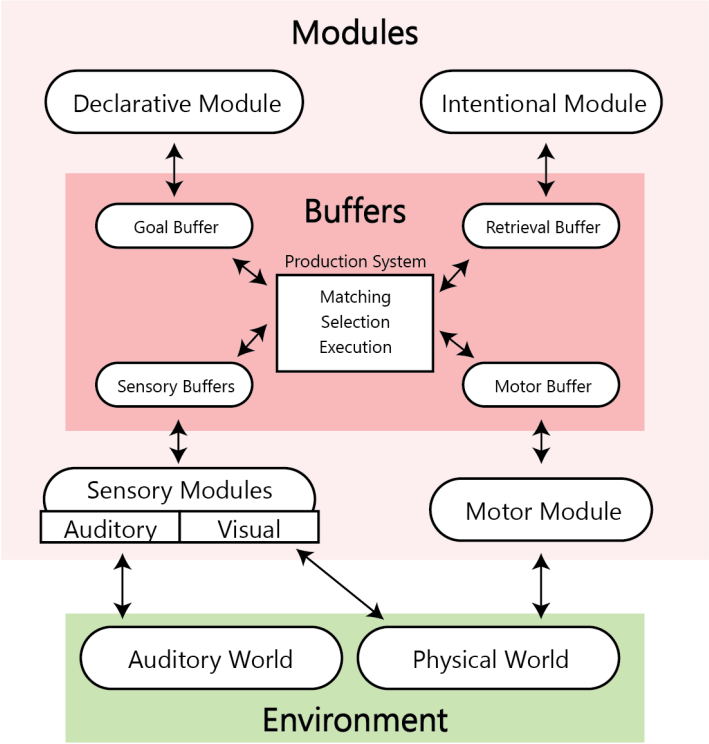

Figure 3.2: A simplified flowchart of the ACT-R model architecture.[3]

3.5 Embodied Cognition

As mentioned in Chapter 1, Saty Raghavachary emphasises the need for models to directly experience the world through the physical body in order to generate human-level grounded meaning [Rag24]. This has been missing from the architectures we have discussed so far, so combining them with the robotic technologies discussed in Section 3.4 would not give models access to some of the key benefits of embodiment. While all of the models from the previous section have contained modules from receiving sense data and making inferences, this has either involved updating an internal model of the world, updating procedural rules of action, or quickly activating a reflexive response. By contrast, Raghavachary argues that agents should associate symbols with real-world experiences in order to ground their meaning in reality[4]. In addition, we have discussed how Andy Clark argues that being able

[3]Image Source: Wikimedia Commons
[4]This is also supported by [Bav+25].

to associate particular motor routines and visual cues with particular data structures allows agents to deliberate more efficiently and rely less heavily on a complex system of internal representations [Cla98].

Furthermore, while the cognitive models described in the previous section represent a reasonably high level of agency and the ability to act within the world by demonstrating the ability to perceive, plan, act, and learn, these architectures place little emphasis on the ability to explore and learn through real-time interactions and experiments with external objects and other agents. This ability is the key to why embodied agency has the potential to be so beneficial for AI models: embodied agents can act with the intention of finding out more about the physical world. We might say this is more like the way that a baby learns – babies explore the world primarily through sensory exploration without much in the way of preexisting procedures for action. The trial and error that babies undergo allows for the refinement of both knowledge and motor functions themselves. For this reason, we now turn to some model architectures that emphasise the learning of inferences through sensory grounding.

3.5.1 Sensory Grounding

The *Connectionist Learning with Adaptive Rule Induction On-Line* (CLARION) Architecture (Fig.3.3) is a dual-process cognitive architecture that makes a distinction between implicit and explicit learning [Sun16]. While its embodiment has only been simulated so far, its structure centres on learning from experience and the consequences of action, while also using declarative knowledge and beliefs to inform action (much like the agentive feedback loop). CLARION is designed to represent embodied cognition and the idea that thinking is deeply influenced by interaction with the external world, a claim which I have also made through my discussion of embodied cognition, grounded meaning, and the *Extended Mind Hypothesis*. CLARION is also designed to be integrated with robotic components such as sensors and 'limbs', although this has not yet been attempted in earnest.

Crucially, CLARION contains both an action-centred and a non-action-centred subsystem. At the bottom level of the action subsystem, a network of perceptrons learns via reinforcement, whereas, at the top level, explicit symbolic rules are extracted and refined. These rules pertain to what action should be taken given a certain internal state, which means that the model carries out explicit reasoning for action selection based on the expected future reward. Both layers in the action-centred subsystem control internal and external actions. At the bottom level of the non-action subsystem, neural networks learn via association, and, at the top level, explicit rules are extracted. These rules are either semantic (pertaining to definitions) or episodic (pertaining to sequences of events) and are grounded in first-hand sensorimotor traces. This means that the model's understanding of meaning and sequences of events are grounded in the sense data received and motor signals employed when learning these rules. In both subsystems, the two layers work together in that repeated successful neural activations that are learnt at the bottom level are extracted to generate new explicit rules or adapt existing rules. These explicit rules then bias or shape further implicit learning at the lower level, meaning that hypotheses can be tested.

Further still, CLARION contains a motivational subsystem, with the bottom level representing simple drives such as hunger and the top level representing explicit goals. The final subsystem is metacognitive and balances the layers of the other subsystems, such as weighting explicit versus implicit responses.

CLARION is similar to ACT-R in that the distinction between action-centred and non-action-centred

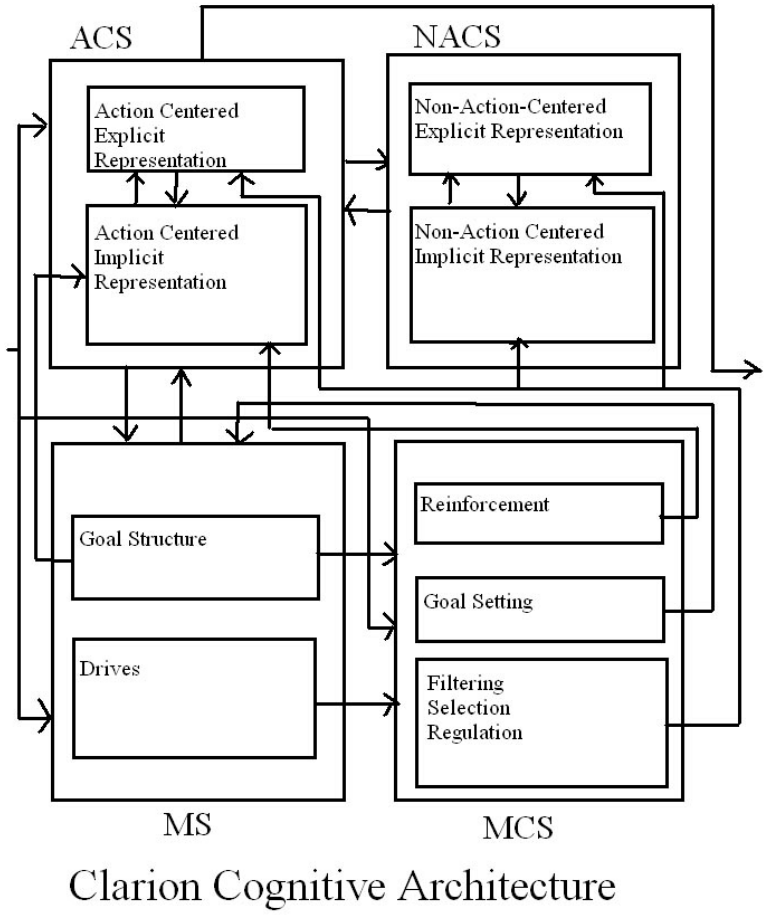

Figure 3.3: A simplified flowchart of the CLARION model architecture.[5]

subsystems is similar to ACT-R's distinction between procedural and declarative memory. However, ACT-R does not distinguish between implicit and explicit processes. Furthermore, rules learnt by ACT-R are not grounded in sensorimotor data as they are in CLARION.

CLARION contains many of the features of high-level agency that I have discussed: reacting in real-time to changes in the environment, learning explicit rules from experience, balancing simpler desires and long-term goals, choosing actions based on expected reward, and grounding meaning in sensorimotor feedback. By building on the research that produced the models discussed throughout this chapter, CLARION has the potential, when combined with devices for sensing and acting within the real physical world, to be a high-level embodied agent. Such an agent, with it's capabilities, has the potential to overcome a great deal of the current limitations of AI models. Of course, with grounded meanings (and, potentially, with multimodal, continuous sense data), CLARION may be more data-efficient than many current modes. As it learns explicit rules through experience, when it repeatedly observes sequences of events, it is able to form if-then rules in the explicit layers, representing causal understanding. Further still, CLARION monitors whether these rules continue to produce expected results, modifying or discarding them if they do not, ensuring that coincidence or noise can be dealt with. CLARION is also able to generalise patterns in the implicit layer from one domain to

[5]Image Source: Wikimedia Commons

another, and adapt explicit rules across contexts, allowing for more efficient transfer learning. Finally, CLARION contains a metacognitive subsystem that monitors, directs, and modifies the operations of all other subsystems, ensuring that processes run smoothly and efficiently. This kind of metacognitive capability lays the groundwork for increasingly self-aware AI systems that can check and explain their reasoning and decisions.

As a result, the CLARION architecture system is able to model a wide range of cognitive abilities, such as reaction time, grammar learning, and complex problem solving. It has been shown to be useful in social situations, in collaborating with humans, and in making decisions in complex, dynamic environments. These capabilities further reinforce how agentive architectures can help AI models to overcome their limitations. In particular, being able to act within an environment and ground learning in sensorimotor feedback is part of what allows models to achieve a more general intelligence, being able to explore and adapt to new scenarios and learn explicit rules through experience. This ability sets CLARION apart for the other architectures I have discussed. Though it naturally has its own limitations, the discussion so far demonstrates that agency and embodiment will be fundamental in the push towards general artificial intelligence.

3.5.2 Robots

We now turn to a way in which key ideas from the model architectures we have discussed can be integrated into a robotic system. The *Distributed Integrated Affect, Reflection, and Cognition* (DIARC) system is a robotic cognitive architecture that integrates modules for perception, natural language processing and generation, procedural and declarative memory, action execution and planning, motivation, and performance monitoring (metacognition) [Sch+19]. New concepts can be added via natural language instruction, which distinguishes it from the models we have looked at so far, as none of these have specific natural language processing frameworks. In addition, DIARC adapts its attention and learning priorities based on sensory data and certainty levels. For example, a surprised human face may lead to focussing on nearby objects. However, DIARC is an architecture that focusses primarily on multimodal human-robot interaction, so it lacks the framework for more explicit symbolic reasoning and complex decision-making.

While DIARC is not designed for a specific robot, it demonstrates the ways in which a robot might be able to act as a being in the world and learn not only from sense data but also from interaction with other agents. If some combination of CLARION's human-like learning and DIARC's interaction capabilities is possible in the future, then we may be able to fully represent the ways in which high-level agents can learn from hypothesis testing in the external world and from interaction with other agents; reason implicitly and explicitly; and react quickly in pursuit of both short-term needs and long-term goals.

3.5.3 Additional Components

We almost have an overview of how agency may be realised in artificial intelligence architectures, based on current frameworks. There are some final components that are worth mentioning which may help to further strengthen agency in AI models.

The first is curriculum-driven embodied learning. This is when agents are guided, whether by another

agent or a specific 'curriculum scheduler' component, through a structured learning process through progressively more complex experiences [XLZ25] [Ram+23]. For example, an infant learns about tools by first mastering touch, then grasping and moving in various ways, then picking up and using tools, and finally grounding this in language. This is similar to how implicit learning can feed explicit learning, except that there is an agent or module in place to ensure that more basic and general skills are mastered first before higher-level skills can be worked on. A curriculum scheduler component in AI organises and adjusts the order and difficulty of the learning tasks that the agent will go through based on their current capabilities and progress. This system has the potential to aid in the transferability of the skills the agent learns, as they are able to master simple control of their motor functions and basic understanding of sensory data before learning more complex solutions to tasks, meaning that agents are able to approach new tasks with a preexisting set of skills for exploration. In addition, this helps to prevent catastrophic forgetting, as earlier tasks are periodically revisited.

Secondly, there may be some additional components required for social grounding. Being able to model the reasoning processes of other agents is not something architectures like CLARION are specifically designed to do. However, InteRRaP does have components for this, so there is a possibility for these elements to be combined for models to be better social agents.

Finally, although the more complex models we have discussed generate expectations by storing explicit rules about the world and planning actions with specific goals in mind, they do not necessarily explicitly simulate the expected world-state that will result from an action. The integration of an imagination component, such as that employed by the *Imagination-Augmented Argent* architecture developed by DeepMind, would allow agents to simulate different actions when deliberating as to what to do [Web+17]. Being able to do so has the potential to enhance action efficacy, as some experimentation is able to be simulated internally. However, more testing is needed to determine whether these benefits could be encapsulated by the explicit reasoning already discussed.

3.6 Future Developments

Given the capabilities of agentive models like CLARION, robotic architectures like DIARC, robot mechanics like those of Atlas, and other components, we now have a clear roadmap for how AI models could be high level embodied agents. As discussed, CLARION, already performs agency at a high level and, if integrated with sensing devices and robotic components, would be considered a high level embodied agent. It interprets sense data it receives to learn patterns and explicit rules, using these to determine plans for the pursuit of immediate needs and long-term goals based on expected rewards, and learns from the outcomes of its actions, thereby refining its reasoning processes and expectation functions.

With its emphasis on exploring the environment and learning through experience, it is able to develop better causal understanding and transfer learned solutions to new domains much better than current AI models that act within limited virtual worlds or models that cannot learn from experience. Further still, employing both explicit and implicit learning aligns much more closely with human cognitive development and improves data efficiency in problem solving and decision-making. CLARION also has the metacognitive ability to reflect upon and refine the performance of all of its submodules, potentially paving the way for more transparent AI systems.

What Clarion lacks, however, is much of a framework for interacting with other agents. This is where DIARC comes in. The ability to learn from natural language instructions, as well as model other agents (a feature that also came up in my discussion of InteRRaP), further enhances how much agents can learn from and control their external environments. Collaboration between agents is an important way that knowledge and resources are shared, and understanding collaborators as agents themselves has the potential to lead to enhanced performance in social reasoning models. Social and group agency is unfortunately beyond the scope of this thesis, but is an interesting area for further research.

Finally, these capabilities could be further enhanced if there is some oversight within the model that dictates the order and complexity of learning tasks. Implementing a component like this could further reduce learning times. In addition, there is also research being done into more detailed simulation of actions (which we might call imagination), that may yet further enable models to make accurate action decisions in complex situations by comparing simulated sequences of events, as opposed to relying simply on a reward function.

If these architectures (CLARION, DIARC, a curriculum scheduler, and an imaginative architecture) could in the future be combined into one architecture and given control over high quality sensing devices (such as cameras and microphones) and mechanical object manipulation tools (such as those akin to limbs and hands), then the resulting intelligent robot would be a high level embodied agent with the potential to develop a detailed understanding of the physical world and the laws that govern it. Such an agent could learn through both experience and from natural language inputs; reflect upon, explain, and refine its own reasoning; solve complex problems across a range of domains including in hypothetical scenarios; and react in real-time to changes in the environment whilst also pursing and updating its long-term goals, to name a few possibilities. These abilities are some of the fundamental expectations that we have for an Artificial General Intelligence and, if achieved, would certainly represent a significant step towards that goal[6].

Although the separate architectures and components that could be brought together to achieve this do exist, it remains to be seen if it is feasible for them all to be brought together. Both CLARION and DIARC are designed to be integrated with robotic components, so this part is not an issue, but whether the two of them could be combined is unclear. DIARC has previously been integrated with ACT-R (and models like it) and found to have the capabilities of both architectures as well as new capabilities that resulted from both systems leveraging one another [SHS13]. This is evidence that the same may be possible with CLARION, although potentially more challenging due to the model's complexity. Further still, CLARION's modular design makes it a good candidate for integrations with other components, such as a curriculum scheduler or imaginative architecture. It is therefore reasonable to suggest that some form of embodied artificial agent is highly possible in the future, and that said agent would have capabilities that overcome many of AI's current limitations.

3.7 Conclusions

In this chapter, I have examined all the ways in which embodied agency could be developed in artificial intelligence models. At the rudimentary level, models can be equipped with some simple sensing devices

[6]I do not claim that this is all that would be needed for AGI, and whether or not true general intelligence can be achieved by AI models is outside the scope of this thesis.

and reaction architectures, as well as possibly also an attention framework for physically or intentionally directing the senses for more efficient reasoning. While these models would not be considered agents, I have shown that the addition of multimodal and continuous sense data as well as the ability to focus on particular stimuli has significant performance benefits for models and that sense devices and attentional frameworks would be necessary for a model that is an embodied agent.

At a middling level, deliberative architectures use beliefs about the world, desires for future world-states, to generate intentions for actions. These models are therefore low-level agents. Deliberative architectures are poor at responding to immediate changes in the environment, so can be combined with reactive architectures to balance more complex reasoning with reflexive reactions to the world. The resulting hybrid architectures are better at responding in real time to changes in the environment, while maintaining the ability to carry out symbolic reasoning.

Being able to update action procedures from experience pushes agency to a higher level. This can be found in cognitive architectures such as ACT-R and further strengthened by the ability of architectures such as CLARION to ground learning and stored rules in sensorimotor experience, which is much closer to the way humans learn about the world. This capability helps to improve cognitive efficiency and further demonstrates the symbiotic relationship between agents and the worlds they inhabit.

Finally, researchers are beginning to develop robotic architectures such as DIARC that can learn from direct communication with human agents, as well as components that plan learning tasks for increased efficiency and imaginative architectures that can simulate possible sequences of events. In the future, if these high-level capabilities could be integrated into one architecture and given control over robotic devices for sensing and manipulating the physical world, then models may be able to achieve human-like embodied agency and reach higher levels of general intelligence. There is already mounting evidence that this kind of integration between architectures and robotic devices is possible and, if achieved, the resulting model would be able to overcome the limitations identified in Chapter 1.

Chapter 4

Conclusions

In this thesis, I have aimed to argue that the some of the limitations faced by current AI models can be overcome through embodied agency.

In Chapter 1, I identified the key limitations faced by current AI models, including a lack of data efficiency, difficulties with causal reasoning, poor transfer learning, and a lack of metacognition. I argued that, while some models do fulfil the basic requirements for agency, they are low-level agents, as agency exists on a spectrum. Higher level agents are better at pursuing a range of short and long-term goals by making accurate predictions about the consequences of their actions and updating their decision-making processes based on the sense feedback that results from those actions. Furthermore, I argued that some AI agents are only able to sense and act within limited virtual worlds (such as the world of the game Go), meaning they are limited in the range and complexity of the problems they can solve. Embodied agency gives models access to the same physical world that humans can learn from and interact with. This allows them to solve complex problems about, and react in real time to the physical world. A higher level of agency can therefore be achieved through embodiment as agents have direct access to much more sense data and can carry out experiments within the physical environment. Finally, I argued that a high level of embodied agency can help models to overcome many of their current limitations. Data efficiency is improved when models are able to ground meaning in the external world. Causal reasoning is improved when models can learn from repeated observations and test hypotheses. Transfer learning, adversarial attacks, and reward hacking can be improved by giving models access to information about a broad range of domains and wider context. A high level of agency may also help models to develop a level of self-awareness, as they can reflect upon and improve their decision-making processes. Finally, a high level of embodied agency allows models to understand themselves as beings within the world and potentially to understand and model the behaviours of other agents.

In Chapter 2, I explored the definitions of embodiment and agency in more detail, making the case that agency requires a sense of embodiment. I argued that there are three requirement for embodiment: being a collection of matter that exists in space and time (The Physical Requirement), being able to sense the external world (The Sense Requirement), and being able to directly control the movement of some objects in the external world (The Manipulation Requirement). I also argued that the parts of the body need not be contiguous with the matter where thinking takes place. I then argued that

agency requires the ability to act intentionally, with an expectation of the outcome of the action, as a causal link with intentional mental states alone can lead to the misattribution of intention. I explored two evolutionary explanations of how organisms developed the ability to set expectations and refine that ability by learning from experience. This sensorimotor feedback loop is essential for high levels agency, allowing models to get better at choosing actions that align with their goals. Finally, I argued that, while agents do not need to be able to act in the physical world, they do need to be able to act in an external world, meaning that only being able to perform mental actions is not enough for agency. For a similar reason, the ability receive sense feedback is an essential part of agency, as agents must be able to compare the outcomes of their actions with expectations. Therefore, a thinking thing must have a sense of embodiment within the external world it has access to in order to be an agent. Embodied agency provides a level of control over the physical world.

In Chapter 3, I looked at the technologies and model architectures that could be used for an AI model to achieve embodied agency. I showed that access to continuous, multimodal sense data has significant performance benefits for models. Furthermore, attentional models can allow for the directing of these senses towards significant stimuli. While these elements are not sufficient for agency, they do lay the groundwork for embodiment. I then followed a timeline of the development of agentive architectures, beginning with deliberative architectures, followed by reactive and hybrid architectures, before finally exploring cognitive architectures and architectures designed for embodied cognition. I made the argument that some combination of the CLARION and DIARC architectures, along with some curriculum scheduling and imaginative components, and combined with robotic devices for sensing and manipulating the physical world, not only meets all the requirements for a high level of embodied agency, but also has the potential to entirely overcome many of the major limitations of modern AI systems. I argued that the integration of these systems is possible in the future, as similar systems have already been integrated and shown to have capabilities beyond the individual capabilities of the systems. There is therefore a clear pathway for the creation of an embodied, high-level artificial agent that is capable of complex reasoning over multiple domains, learning implicitly and explicitly from experience and from natural-language inputs, communicating with and modelling other agents, building cognitive skills through experimentation in a similar way to humans, imagining a range of hypothetical situations, and pursuing complex long and short-term goals. This would represent a significant step in the push towards general intelligence.

In this thesis, I have made the case that a high level of agency is vital to the way that humans are able to develop their cognitive abilities and understanding of the world. It is not just the individual skills that make up agency (interpreting sense data, action selection, expectation generation, updating procedures from sensorimotor feedback) that aid in the cognitive development of humans, but agency as a whole, which leads humans to understand ourselves as beings within the world, within which we have the ability to experiment and pursue a range of goals. This exploration is made possible because we are embodied creatures, able to directly sense and manipulate the objects around us. I have argued that current AI models could see significant performance benefits via high level embodied agency, particularly through grounding knowledge in external experience, experimenting across a range of domains, and constantly reviewing and refining internal processes. I acknowledge that there are potentially less computationally taxing solutions to some of the individual problems faced by modern AI, but argued that high level embodied agency has the potential to overcome a high number of problems,

while significantly boosting performance. While high levels of agency are possible without embodiment, models would then be limited by the complexity of the simulated world that they have access to, and simulating a world of a similar complexity to our own requires high levels of computational power. I have argued that the technological groundwork has already been laid for the creation of high-level embodied agents, and that doing so would be a significant step not just towards more powerful and efficient models, but towards some kind of artificial general intelligence.

In future, I would like to be able to explore some of the topics that were outside the scope of this thesis, such as mental actions, group agency and emotions. I would also like to delve deeper in to the precise ways that our existing models for agency and embodied cognition could be integrated into one system.

Bibliography

[Ada21] Amina Adadi. "A Survey on Data-Efficient Algorithms in the Big Data Era". In: *Journal of Big Data* 8.1 (2021), page 24. https://doi.org/10.1186/s40537-021-00419-9 (cited on page 7).

[Ara+25] Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R. Glass, and Hilde Kuehne. *CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment*. 2025. https://arxiv.org/abs/2505.01237 (cited on page 34).

[Ari99] Aristotle. *Nicomachean Ethics*. Edited by Terence Irwin. 2nd. Indianapolis: Hackett Publishing, 1999 (cited on page 3).

[Arm68] D. M. Armstrong. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul, 1968. ISBN: 9780203003237 (cited on page 28).

[Avi06] Avicenna. *A Compendium on the Soul*. Translated by Edward Abbott van Dyck. Originally written in the 11th century; English translation from the Arabic. Verona, Italy: Stamp. di Nicola Paderno, 1906 (cited on page 19).

[Ban89] Albert Bandura. "Human Agency in Social Cognitive Theory". In: *American Psychologist* 44.9 (1989), pages 1175–1184. https://doi.org/10.1037/0003-066X.44.9.1175 (cited on page 3).

[Bav+25] Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. "Modelling Multimodal Integration in Human Concept Processing with Vision-Language Models". In: *ArXiv* (2025). preprints. https://arxiv.org/abs/2407.17914 (cited on pages 33, 43).

[BIP88] Michael E. Bratman, David J. Israel, and Martha E. Pollack. "Plans and resource-bounded practical reasoning". In: *Computational Intelligence* 4.3 (1988), pages 349–355. https://doi.org/10.1111/j.1467-8640.1988.tb00284.x (cited on page 40).

[Bro86] R. Brooks. "A robust layered control system for a mobile robot". In: *IEEE Journal on Robotics and Automation* 2.1 (1986), pages 14–23. https://doi.org/10.1109/JRA.1986.1087032 (cited on page 41).

[Buc23] Cameron J. Buckner. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. 1. Oxford University Press, 2023. ISBN: 9780197653302 (cited on pages 6, 7, 35).

[CA16] Jack Clark and Dario Amodei. *Faulty reward functions in the wild*. https://openai.com/index/faulty-reward-functions/. 2016 (cited on page 7).

[Car+24] João Carreira, Michael King, Viorica Pătrăucean, Dilara Gokay, Cătălin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, Dima Damen, and Andrew Zisserman. "Learning from One Continuous Video Stream". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. https://doi.org/10.48550/arXiv.2312.00598. https://arxiv.org/abs/2312.00598 (cited on pages 14, 33).

[CC98] Andy Clark and David J. Chalmers. "The Extended Mind". In: *Analysis* 58.1 (1998), pages 7–19 (cited on page 22).

[Cha23] David J. Chalmers. "Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models". In: *Proceedings and Addresses of the American Philosophical Association* 97 (2023), pages 22–45 (cited on page 19).

[Cla98] Andy Clark. "Embodiment and the Philosophy of Mind". In: *Royal Institute of Philosophy Supplement* 43 (1998), pages 35–51. https://doi.org/10.1017/S135824610000429X (cited on pages 14, 44).

[Cou24] CourseMentor. *200+ Robotics Research Topics: Discovering Tomorrow's Tech*. 2024. https://coursementor.com/blog/robotics-research-topics/ (cited on page 32).

[CWF24] Christine Cuskley, Rebecca Woods, and Molly Flaherty. "The Limitations of Large Language Models for Understanding Human Language and Cognition". In: *Open Mind* 8 (Aug. 2024), pages 1058–1083. ISSN: 2470-2986. https://doi.org/10.1162/opmi_a_00160 (cited on page 4).

[Dav01] Donald Davidson. "Agency". In: *Essays on Actions and Events*. Oxford: Oxford University Press, 2001, pages 43–62 (cited on page 3).

[Dav63] Donald Davidson. "Actions, Reasons, and Causes". In: *The Journal of Philosophy* 60.23 (1963), pages 685–700. https://doi.org/10.2307/2023177 (cited on page 23).

[Dav73] Donald Davidson. "Freedom to Act". In: *Essays on Actions and Events*. Oxford University Press, 1973, pages 79–92 (cited on page 23).

[DeL11] Manuel DeLanda. *Philosophy and Simulation: The Emergence of Synthetic Reason*. London: Continuum, 2011. ISBN: 9781441170286 (cited on page 3).

[Des08] René Descartes. *Meditations on First Philosophy: With Selections from the Objections and Replies*. Translated by Michael Moriarty. Originally published in 1641. Oxford: Oxford University Press, 2008 (cited on page 18).

[DS03] Chris Dobbyn and Susan Stuart. "The Self as an Embedded Agent". In: *Minds and Machines* 13.2 (2003), pages 187–201. https://doi.org/10.1023/A:1022997315561 (cited on pages 11, 16).

[Fer92] Innes A. Ferguson. "Touring Machines: Autonomous Agents with Attitudes". In: *Computer* 25.5 (May 1992), pages 51–55. ISSN: 0018-9162. https://doi.org/10.1109/2.144395 (cited on page 41).

[FN71] Richard E. Fikes and Nils J. Nilsson. "Strips: A new approach to the application of theorem proving to problem solving". In: *Artificial Intelligence* 2.3 (1971), pages 189–208. ISSN: 0004-3702. https://doi.org/10.1016/0004-3702(71)90010-5 (cited on page 39).

[Gat92] Erann Gat. "Integrating planning and reacting in a heterogeneous asynchronous architecture for controlling real-world mobile robots". In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI'92. San Jose, California: AAAI Press, 1992, pages 809–815. ISBN: 0262510634 (cited on page 41).

[Gau09] Mary Gauvain. "The Stages of the Intellectual Development of the Child". In: *Readings on the Development of Children*. Edited by Mary Gauvain and Michael Cole. New York: Worth Publishers, 2009 (cited on page 4).

[GG18] Shani Gamrian and Yoav Goldberg. "Transfer Learning for Related Reinforcement Learning Tasks via Image-to-Image Translation". In: *arXiv preprint arXiv:1806.07377* (2018) (cited on page 8).

[GI89] Michael P. Georgeff and Francois Felix Ingrand. "Decision-making in an embedded reasoning system". In: *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'89. Detroit, Michigan: Morgan Kaufmann Publishers Inc., 1989, pages 972–978 (cited on pages 39, 40).

[God16] Peter Godfrey-Smith. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Farrar, Straus and Giroux, 2016. ISBN: 9780374227760 (cited on page 24).

[Gol06] Alvin I. Goldman. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford, UK: Oxford University Press, 2006 (cited on page 27).

[Gon+23] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. *Contrastive Audio-Visual Masked Autoencoder*. 2023. https://arxiv.org/abs/2210.07839 (cited on page 34).

[Gru04] Rick Grush. "The Emulation Theory of Representation: Motor Control, Imagery, and Perception". In: *Behavioral and Brain Sciences* 27.3 (2004), pages 377–396. https://doi.org/10.1017/S0140525X04000093 (cited on page 25).

[GSS15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *arXiv preprint arXiv:1412.6572* (2015) (cited on page 7).

[Hat14] J. H. van Hateren. "The Origin of Agency, Consciousness, and Free Will". In: *Phenomenology and the Cognitive Sciences* 14.4 (2014), pages 979–1000. https://doi.org/10.1007/s11097-014-9396-5 (cited on page 26).

[Hay81] Patrick J. Hayes. "The Frame Problem and Related Problems in Artificial Intelligence". In: *Readings in Artificial Intelligence*. Edited by Bonnie Lynn Webber and Nils J. Nilsson. Morgan Kaufmann, 1981, pages 223–230. ISBN: 978-0-934613-03-3. https://doi.org/10.1016/B978-0-934613-03-3.50020-9 (cited on page 40).

[Him16] Johannes Himmelreich. "Agency and Embodiment: Groups, Human–Machine Interactions, and Virtual Realities". In: *Ratio* 31.2 (2016), pages 197–213. https://doi.org/10.1111/rati.12158 (cited on page 28).

[Hua+25] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2 (2025), pages 1–55. ISSN: 1558-2868. https://doi.org/10.1145/3703155 (cited on page 9).

[Hue21] Charlie Huenemann. "Hobbes, Automata, and Autonomy". In: *Clio: A Journal of Literature, History, and the Philosophy of History* 48.3 (2021), pages 325–342. https://www.academia.edu/49494861/Hobbes_automata_and_autonomy (cited on page 3).

[Hum07] David Hume. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford: Oxford University Press, 2007 (cited on page 3).

[Hus52] Edmund Husserl. *Ideas: General Introduction to Pure Phenomenology*. Originally published in 1913. New York: Macmillan, 1952 (cited on page 19).

[Inf25] Aaron Infante. "Machine Learning vs Rule-Based Systems: Key Differences Explained". In: *Greenbot* (2025). https://www.greenbot.com/machine-learning-vs-rule-based/ (cited on page 6).

[Kah11] Daniel Kahneman. *Thinking, Fast and Slow*. New York: New York: Penguin Books, 2011 (cited on page 41).

[Kui+15] Scott Kuindersma, Robin Deits, Maurice Fallon, Andrés Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. "Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot". In: *Autonomous Robots* 40 (July 2015). https://doi.org/10.1007/s10514-015-9479-3 (cited on page 37).

[Lee00] Jaeho Lee. "Reactive-System Approaches to Agent Architectures". In: *Intelligent Agents VI. Agent Theories, Architectures, and Languages*. Edited by Nicholas R. Jennings and Yves Lespérance. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pages 132–146. ISBN: 978-3-540-46467-9 (cited on page 41).

[Li+25] Xiaojie Li, Yibo Yang, Jianlong Wu, Jie Liu, Yue Yu, Liqiang Nie, and Min Zhang. *Continuous Knowledge-Preserving Decomposition for Few-Shot Continual Learning*. 2025. https://arxiv.org/abs/2501.05017 (cited on page 38).

[Lin+24] Guanyu Lin, Tao Feng, Pengrui Han, Ge Liu, and Jiaxuan You. "Paper Copilot: A Self-Evolving and Efficient LLM System for Personalized Academic Assistance". In: *arXiv preprint arXiv:2409.04593* (2024). https://arxiv.org/abs/2409.04593 (cited on page 7).

[Lis21] Christian List. "Group Agency and Artificial Intelligence". In: *Philosophy & Technology* 34 (2021), pages 1213–1242. https://doi.org/10.1007/s13347-021-00454-7 (cited on page 3).

[Mer12] Maurice Merleau-Ponty. *Phenomenology of Perception*. Translated by Donald A. Landes. Originally published in 1945 as *Phénoménologie de la perception*. London: Routledge, 2012. ISBN: 9780415834339 (cited on page 19).

[MH69] John McCarthy and Patrick J. Hayes. “Some Philosophical Problems from the Standpoint of Artificial Intelligence”. In: *Machine Intelligence 4*. Edited by B. Meltzer and D. Michie. Edinburgh University Press, 1969, pages 463–502 (cited on page 40).

[MN23] Lauren G. Malachowski and Amy Work Needham. “Chapter Two - Infants exploring objects: A cascades perspective”. In: *Developmental Cascades*. Edited by Catherine S. Tamis-Lemonda and Jeffrey J. Lockman. Volume 64. Advances in Child Development and Behavior. JAI, 2023, pages 39–68. https://doi.org/10.1016/bs.acdb.2022.11.001 (cited on page 7).

[MP94] Jörg P. Müller and Michael Pischel. “The INTERRAP Agent Architecture”. In: *Proceedings of the Second International Conference on Autonomous Agents*. New York, NY, USA: ACM, 1994, pages 376–377 (cited on page 41).

[Nei+96] Ulric Neisser, Abid Boodoo, Thomas J. Bouchard Jr, A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert J. Sternberg, and Susana Urbina. *Intelligence: Knowns and Unknowns*. Washington, DC: American Psychological Association, 1996 (cited on page 4).

[PC09] Giovanni Pezzulo and Cristiano Castelfranchi. “Thinking as the Control of Imagination: A Conceptual Framework for Goal-Directed Systems”. In: *Psychological Research* 73.4 (2009), pages 559–577. https://doi.org/10.1007/s00426-009-0237-z (cited on page 26).

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. Cambridge University Press, 2009 (cited on page 8).

[Pez11] Giovanni Pezzulo. “Grounding Procedural and Declarative Knowledge in Sensorimotor Anticipation”. In: *Mind & Language* 26.1 (2011), pages 78–114. https://doi.org/10.1111/j.1468-0017.2010.01411.x (cited on page 26).

[PT23] Juan S. Piñeros Glasscock and Sergio Tenenbaum. “Action”. In: *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University, 2023 (cited on page 23).

[Put81] Hilary Putnam. *Reason, Truth and History*. Cambridge: Cambridge University Press, 1981. ISBN: 9780521297752 (cited on page 29).

[Rag24] S. Raghavachary. “The Embodied Intelligent Elephant in the Room”. In: *Proceedings from Springer Nature*. Springer Nature, 2024, pages 716–722. https://doi.org/10.1007/978-3-031-50381-8_77 (cited on pages 14, 43).

[Ram+23] Ram Ramrakhya, Dhruv Batra, Aniruddha Kembhavi, and Luca Weihs. “Curriculum Learning via Task Selection for Embodied Navigation”. In: *Embodied AI Workshop at CVPR*. 2023. https://embodied-ai.org/papers/2023/22.pdf (cited on page 47).

[RG91] Anand S. Rao and Michael P. Georgeff. “Modeling Rational Agents within a BDI-Architecture”. In: *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*. Edited by James Allen, Richard Fikes, and Erik Sandewall. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991, pages 473–484. http://jmvidal.cse.sc.edu/library/rao91a.pdf (cited on page 39).

[Rob24] RoboticsBiz. *500 Recent Research Papers and Projects in Robotics*. 2024. https://roboticsbiz.com/500-recent-research-papers-and-projects-in-robotics-free-download/ (cited on page 32).

[RTO18] Frank Ritter, Farnaz Tehranchi, and Jacob Oury. "ACT-R: A cognitive architecture for modeling cognition". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 10 (Dec. 2018), e1488. https://doi.org/10.1002/wcs.1488 (cited on page 42).

[Sch+19] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. "An Overview of the Distributed Integrated Cognition Affect and Reflection DIARC Architecture". In: *Cognitive Architectures*. Edited by Maria Isabel Aldinhas Ferreira, João Silva Sequeira, and Rodrigo Ventura. Cham: Springer International Publishing, 2019, pages 165–193. ISBN: 978-3-319-97550-4. https://doi.org/10.1007/978-3-319-97550-4_11 (cited on page 46).

[Sch+23] Laura Schnetzer, Mark McCoy, Jürgen Bergmann, Alexander Kunz, Stefan Leis, and Eugen Trinka. *Locked-in syndrome revisited*. 2023. https://doi.org/10.1177/17562864231160873 (cited on page 20).

[Sch19] Markus Schlosser. "Agency". In: *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University, 2019 (cited on page 22).

[SH23] Arlette Streri and Maria Dolores de Hevia. "How do human newborns come to understand the multimodal environment?" In: *Psychonomic Bulletin & Review* 30 (2023), pages 1171–1186. https://doi.org/10.3758/s13423-023-02260-y. https://link.springer.com/article/10.3758/s13423-023-02260-y (cited on page 14).

[SHS13] Matthias Scheutz, Jack Harris, and Paul Schermerhorn. "Systematic Integration of Cognitive and Robotic Architectures". In: *Advances in Cognitive Systems* (2013), pages 277–296 (cited on page 48).

[SLS24] Kun Su, Xiulong Liu, and Eli Shlizerman. *From Vision to Audio and Beyond: A Unified Model for Audio-Visual Representation and Generation*. 2024. https://arxiv.org/abs/2409.19132 (cited on page 34).

[SS24] Gaurav Shrivastava and Abhinav Shrivastava. *Continuous Video Process: Modeling Videos as Continuous Multi-Dimensional Processes for Video Prediction*. 2024. https://arxiv.org/abs/2412.04929 (cited on page 33).

[Sun16] Ron Sun. *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. Oxford University Press, Mar. 2016. ISBN: 9780199794553. https://doi.org/10.1093/acprof:oso/9780199794553.001.0001 (cited on page 44).

[Vid14] Clément Vidal. "The Future of Scientific Simulations". In: *The Beginning and the End*. The Frontiers Collection. Springer, 2014, pages 153–160. https://doi.org/10.1007/978-3-319-05062-1_7. https://link.springer.com/chapter/10.1007/978-3-319-05062-1_7 (cited on pages 13, 33).

[Wan+16] Fei-Yue Wang, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. *Where Does AlphaGo Go: From Church-Turing Thesis to AlphaGo Thesis and Beyond*. 2016. https://doi.org/10.1109/JAS.2016.7471613 (cited on page 3).

[Web+17] Theophane Weber, Sébastien Racanière, David P. Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, Razvan Pascanu, Peter W. Battaglia, David Silver, and Daan Wierstra. "Imagination-Augmented Agents for Deep Reinforcement Learning". In: *CoRR* abs/1707.06203 (2017). http://arxiv.org/abs/1707.06203 (cited on page 47).

[Weh20] Maren Wehrle. "Being a Body and Having a Body: The Twofold Temporality of Embodied Intentionality". In: *Phenomenology and the Cognitive Sciences* 19.3 (2020), pages 499–521. https://doi.org/10.1007/s11097-019-09610-z (cited on page 20).

[WGJ95] David M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. "An Internal Model for Sensorimotor Integration". In: *Science* 269.5232 (1995), pages 1880–1882. https://doi.org/10.1126/science.7569931 (cited on pages 25, 38).

[Wit53] Ludwig Wittgenstein. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Oxford: Blackwell, 1953 (cited on page 22).

[Woo09] Michael Wooldridge. *An Introduction to MultiAgent Systems*. 2nd. John Wiley & Sons, 2009 (cited on page 42).

[WS24] Mary Jean Walker and Robert Sparrow. "Being in the World: Extended Minds and Extended Bodies". In: *Neuro-ProsthEthics: Ethical Implications of Applied Situated Cognition*. Edited by Jan-Hendrik Heinrichs, Birgit Beck, and Orsolya Friedrich. Berlin, Heidelberg: Springer Berlin Heidelberg, 2024, pages 73–87. ISBN: 978-3-662-68362-0. https://doi.org/10.1007/978-3-662-68362-0_5 (cited on page 22).

[XLZ25] Lipei Xie, Yingxin Li, and Huiping Zhuang. "Analytic Task Scheduler: Recursive Least Squares Based Method for Continual Learning in Embodied Foundation Models". In: *arXiv preprint arXiv:2506.09623* (2025). https://arxiv.org/abs/2506.09623 (cited on page 47).

[ZLD25] Yanbin Zhang, Yang Li, and Zhendong Dai. "Continuous Learning and Adaptation of Neural Control for Proprioceptive Feedback Integration in a Quadruped Robot". In: *Journal of Bionic Engineering* 22.3 (2025), pages 742–765. https://doi.org/10.1007/s42235-025-00742-4. https://link.springer.com/article/10.1007/s42235-025-00742-4 (cited on page 38).