Extended Mind Hypothesis and Extended Knowledge

Francis Payyappilly Jose

A thesis presented for the degree of

Doctor of Philosophy

School of Humanities

College of Arts

University of Glasgow

January 2023

TABLE OF CONTENTS

List of Figures

ACKNOWLEDGMENTS

First of all, I would like to thank my first supervisor, Adam Carter. This thesis would not have been possible without his encouragement and advice, especially during the COVID pandemic. I would also like to thank Christoph Kelp for his support as my second supervisor.

I am grateful to my family, my wife, Digi Francis, and my children Jeevan Francis, Jyothi Francis and Janett Francis for their continued support as I completed this PhD.

I am also indebted to my friends and colleagues, who have been so supportive and cheerful during my research.

1 INTRODUCTION

This thesis mainly focuses on externalism in the philosophy of mind, such as the extended mind (EM) thesis, externalism in epistemology, such as anti-luck virtue epistemology (ALVE), and the ramifications of EM and ALVE for yielding extended knowledge (EK). This thesis also focuses on the fundamental nature of dynamic feedback loops involving a cognitive agent and an external artefact and the formulation of an integrated concept of EM (IEM) that combines dynamical systems theory (DST) as applied to cognition, niche construction theory (NCT), cognitive niches (CNs), cognitive niche construction (CNC), developmental systems theory, and the patterns and history of external artefacts. IEM provides a framework for multidisciplinary integration.

The traditional view of mind, i.e. cognitive internalism, has a Cartesian legacy in which the mind and body are distinct and in which the mind is entirely internal to the agent. In contrast, cognitive externalism claims that some cognitive processes and mental states are not confined internally to the agent, so that the mind has external components. For example, in the EM thesis, a pen and paper used to perform complex calculations may be considered to be external components. In contrast, according to cognitive internalism, the pen and paper have only an add-on or enabling role in performing calculations, meaning that cognition is entirely internal to the agent. Clark & Chalmers's (C&C's) EM thesis claims that an external artefact does play a role in cognition if the external artefact is integrated with the agent in such a way that the external artefact performs some cognitive processes similar to the way that an internal counterpart would. Therefore, cognition can extend into the environment. If cognition extends into the environment, so does the mind. C&C's thesis disagrees with the bio-prejudice that cognition can only be internal to the agent.

Chapter 2: EM thesis and criticisms: This chapter evaluates various criticisms levelled against the EM thesis. Criticisms relating to the enabling versus the constitutive role of an external artefact in cognition and cognitive bloat put some pressure on C&C's version of EM. Clark's reliance on functionalist theory is not sufficient to counter these criticisms.

Chapter 3: Modified EM based on DST: This chapter mainly focuses on the necessary and sufficient conditions for extended cognition (EC) based on DST and is supported by various arguments from developmental systems theory and NCT and from the patterns and historicism embedded in external information-bearing structures and the role of a pattern recogniser. This chapter also explores the integration of EC and NCT to explain how intellectual abilities arise from the innate cognitive abilities of humans endowed by evolution.

Chapter 3 provides insights about (1) modified EM based on DST, (2) the intersection of EC, NCT and CNC, and (3) IEM. In Chapter 3, I propose modified EM based on DST.

I argue that the existence of a feedback loop with an external information-bearing structure and the manipulation of that structure are necessary and sufficient conditions for EM. I have formulated a modified version of Clarkian EM in which the externalist criteria for EM can be explained epistemically and ontologically by feedback loops involving external artefacts. Modified EM addresses the coupling-constitution fallacy. This fallacy arises when there is confusion between merely coupling with an external artefact and that artefact being a constitutive part of a cognitive process. In modified EM, the requirement for active manipulation within a feedback loop clearly delineates when an external artefact transitions from being a mere tool (coupling) to being an integral component of cognition (constitution). Modified EM avoids cognitive bloat by setting a clear criterion for what counts as a constitutive part of cognition: the presence of a feedback loop involving the manipulation of an external information-bearing structure. This criterion is specific and restrictive and prevents the indiscriminate inclusion of external artefacts in cognitive processes. Only those artefacts that are actively manipulated in a feedback loop – and thereby have a direct and significant impact on a cognitive task – are able to extend cognition. The nature of the interaction in the feedback loop is what grants the artefact its constitutive role. By defining specific criteria for cognitive extension, particularly the requirement for a feedback loop involving the manipulation of external artefacts, modified EM effectively addresses the concerns of the coupling-constitution fallacy and cognitive bloat.

Intersection of EC, NCT and CNC: A construal of EC based on DST has the potential for interdisciplinary integration because the central theme of DST-based EC is a dynamic feedback loop between an agent and an artefact. For example, one of the core themes of human NCT is the dynamic interaction of an agent and an artefact. Thus, Chapter 3 also focuses on the application of DST to cognition and provides insights from evolutionary biology through NCT, CN, CNC and developmental systems theory, all of which are generally aligned with the principles of EM. These fields emphasise the importance of both internal and external factors in shaping cognitive processes. Both EC and NCT touch on the idea that human cognition is not solely an internal, brain-based process but also extends into the environment and is shaped by the ways that humans modify their surroundings. This ramification of EC, NCT, CNC and externalism in epistemology provides an interdisciplinary framework that can be applied to multidisciplinary integration.

Integrated EM: The intersections of NCT, EC, CNC and virtue epistemology (VE) provide a theoretical foundation for understanding how humans interact with their environment and how this interaction shapes cognitive processes. Based on this intersection, in Chapter 3, I formulate IEM and explore the versatility and potential of EM based on NCT, DST and developmental systems theories. I develop broad patterns from IEM, which is useful in a wide range of domains and is not limited to being a philosophy of mind.

To explore the potential for the ramifications of DST-based EM with epistemology, the next chapter explores post-Gettier epistemology and how it can assimilate DST-based EC in producing EK. From an epistemological perspective, the proposed version of EM is aligned with EK. It justifies the constitutive role of external artefacts based on feedback loops and the integration of an external artefact with an agent’s cognitive ability that results in an extended cognitive ability that can be used to form beliefs unreflectively. This modified version of EM is immune to criticisms like the coupling-constitution fallacy and cognitive bloat.

Chapter 4: Post-Gettier epistemology, anti-luck epistemology and ALVE: In Chapter 4, I discuss how post-Gettier epistemology is applied to analyse knowledge and explore a possible enrichment of post-Gettier epistemology based on modified EM. The traditional analysis of knowledge, such as a justified true belief (JTB), has a tripartite structure. However, lucky true beliefs, as in Gettier-style cases, can undermine knowledge. This chapter mainly focuses on Gettier cases, various accounts of luck, anti-luck epistemology (ALE), robust virtue epistemology (RVE), ALVE and criticisms of ALVE. The nature and characteristics of luck and its relation to knowledge are still being debated in epistemology. There are various theories of luck, such as probability, the modal account of luck (MAL) and various hybrid accounts, but all these theories of luck have counterexamples and none of them is an adequate theory of luck. However, there is a consensus that knowledge is incompatible with at least some kinds of luck.

Pritchard formulated ALVE by choosing a suitable account of luck that can ensure the safety of a target belief and a virtue theoretic condition to show that cognitive success is due to the cognitive ability of the agent. Pritchard uses MAL to guarantee the safety of a target belief, i.e. that acquired knowledge is not due to luck. Many epistemologists, such as Lackey, Carter, Peterson and De Grefte, have criticised MAL, stating that MAL cannot ensure the full spectrum of luck in knowledge acquisition. Recently, Pritchard moved away from the concept of luck to risk. Pritchard argues that risk is fundamental and forward looking and that luck is backward looking as it is based on what went wrong. Pritchard modified MAL to the modal account of risk (MAR). Accordingly, he then modified his theory of knowledge ALVE to anti-risk virtue epistemology (ARVE).

When an external artefact is involved in knowledge acquisition, MAR alone cannot capture the full spectrum of risk. I propose that to cover the entire spectrum of risk in knowledge acquisition, both probability and MAR are required. Externalism in epistemology, such as ARVE, can accommodate EC. A dynamic feedback loop between an agent and an artefact is essential in EC. However, a risk assessment with an external artefact cannot be completely captured by the modal account in ARVE. Hence, a new account of risk is required. The next chapter explores a potential new account of risk that can accommodate knowledge acquisition involving an external artefact.

Chapter 5: The requirement for a novel account of risk when EC is assimilated into epistemology: Chapter 5 mainly focuses on the potential risks associated with an external artefact during the production of knowledge.

EK scenarios require a risk assessment approach that integrates both the cognitive aspects (how users interact with and interpret information from artefacts) and the technical aspects (reliability and functionality of artefacts). At its core, risk refers to the possibility of something happening that would have a negative impact on objectives, goals, or desired outcomes. Risk, in its most basic form, is about the likelihood and impact of negative events. However, in the context of EK involving artefacts, risk becomes a more complex concept. It necessitates considering not only the traditional aspects of probability and impact but also the broader range of scenarios and implications brought about by the integration of technology into cognitive processes. This comprehensive understanding of risk is crucial for effectively managing potential negative outcomes in both theoretical and practical realms. When discussing risk in the context of EK, where cognitive processes are extended through the integration of artefacts, the concept of risk broadens.

Pritchard’s MAR has limitations as it cannot capture the full spectrum of risk in knowledge production, especially when an artefact is involved, as the risks associated with an artefact in knowledge production require both probabilistic and modal components. A hybrid account of risk covers the full spectrum of risk in knowledge production, as it includes both the modal and probabilistic accounts in a risk assessment.

It is an adequate theory of risk, especially for the risk in knowledge acquisition involving an external artefact, as, roughly, the modal component can address veritic risk and the probability component can address the risks associated with the artefact in knowledge acquisition. In veritic risk, an agent's belief is true in the actual world but false in nearby possible worlds. Thus, the modal and probability components in the hybrid account of risk (HAR) can address the full spectrum of risk associated with knowledge acquisition. HAR is also adequate for risk assessments in engineering.

Pritchard's ARVE has limitations in terms of establishing and differentiating between the relation and interface between an artefact and an agent and the relation and interface between the environment and an agent. This necessitates a modification of ARVE, which I realised by establishing the nature and relationship of the interface between an artefact and an agent and the interface between the environment and an agent.

Modified ARVE (MARVE) is based on HAR, which has probabilistic and modal components that can capture the full spectrum of risks. Thus, HAR can be applied to epistemology and to engineering. This is especially important when an artefact is involved in knowledge acquisition. In engineering, risk is solely based on probability; however, a modal component can capture risk and safety, especially when there is uncertainty in the risk assessment. MARVE integrates feedback loops and artefact manipulation from EC. These elements are crucial in understanding how artefacts contribute to cognitive processes and knowledge acquisition. This integration allows MARVE to evaluate risks associated with artefacts in a more nuanced manner that reflects their role in extended cognitive systems. The next chapter explores EK based on MARVE and its potential merits over Pritchard's construal of EK.

Chapter 6: Pritchard's ARVE, EK, MARVE and modified extended knowledge: Chapter 6 focuses on the ramifications of EM and epistemology in terms of the potential for EK and critically evaluates various criticisms of and debates about EK. The chapter presents two arguments against Pritchard's version of EK. First, it struggles to counter the criticism raised by Adams and Aizawa (A&A) (2001, 2010) and Rupert (2004) regarding the constitutive role of external artefacts instead of their enabling role. The second argument is related to the limits and extent of EK. I also propose a modified version of Pritchard’s EK (MEK) based on MARVE. MEK justifies the constitutive role of external artefacts based on feedback loops and the phenomenological integration of external artefacts with an agent’s cognitive ability to become an extended cognitive ability that can be used to form beliefs unreflectively. Further, Chapter 6 focuses on the various debates around EK, such as the debates between Kelp and Vaesen and between Carter and Jarvis. I argue that Kelp’s version of extended epistemology cannot be consistent with EK, as it fails to counter the internalist criticism relating to the enabling role of external artefacts versus the constitutive role. I claim that the modified versions of ARVE and EK based on the dynamic relationship between an agent and an artefact can dissolve Kelp’s concern about extended agents. In the same way, I show that Carter’s concerns relating to M-parity and E-parity and the problems in assimilating EK with mainstream epistemologies dissolve in MARVE and MEK.

Chapter 7 details the conclusion. The main themes of the thesis are as follows:

(1) DST-based EC to counter criticisms like the coupling-constitution fallacy and cognitive bloat.

(2) The possibility that a potential theory of EC based on DST can lead to the multidisciplinary integration with NCT, CN, CNC and VE and the development of an integrated framework based on IEM that explains the cognitive and intellectual development of humans.

(3) Potential ramifications of modified EM in epistemology and the construal of a new HAR to accommodate the risk associated with an artefact along with addressing veritic luck.

(4) Modification of Pritchard’s ARVE based on HAR and a detailed framework for risk assessment in knowledge acquisition and, finally, a modified account of EK based on MARVE and HAR.

2 Critical Evaluation of the Extended Mind Hypothesis

2.1 Introduction

I asked my daughter: “Where is your mind?” I explained that her mind is composed of her beliefs, desires, hopes, consciousness and experience. She replied that: “My mind is with me, inside my physical boundary.” This is the general perception of mind, i.e. the mind is internal and individuated. However, Clark and Chalmers (C&C) (1998) proposed a different perspective in which the mind can extend into the environment via closely coupled external artefacts that we use to form beliefs. This is the hypothesis of extended mind (EM). An example is the role that pen and paper have in complex computations. A simple multiplication, such as 6 × 7 = 42, can be done in the mind without the help of any external artefact. However, for a complex calculation with various steps, part of the computational process has to be offloaded into one or more external artefacts, such as pen and paper. An example is long multiplication, like 6866 × 8699. In this case, the pen and paper are coupled to the mathematician and their role is essential in the accomplishment of the complex calculation. Clearly, the pen and paper have a part in the cognitive processes. Since the pen and paper have a constitutive role in the overall cognitive process needed to accomplish the calculation, we can say that cognition has extended into the environment.

The traditional view of mind, i.e. cognitive internalism (Carter et al., 2014), has a Cartesian legacy in which the mind and body are distinct and in which the mind is entirely internal to the agent. In contrast, cognitive externalism claims that some cognitive processes and mental states are not confined internally to the agent, so that the mind has external components. For example, in the EM thesis, the external components are the pen and paper used to perform the complex calculation. The EM thesis is compatible with cognitive externalism. In contrast, according to cognitive internalism, the pen and paper have only an add-on or enabling role in performing the calculation, and this thesis asserts that cognition is entirely internal to the agent.

This chapter critically evaluates C&C’s EM hypothesis and explores a way forward for it. My conclusion is that C&C fail to provide necessary and sufficient conditions for the claim that part of the mind is external to the body, i.e. that the mind extends into the environment via external artefacts. However, I am going to argue in Chapter 3 that a modified version of Clark’s EM based on dynamic feedback loops between a cogniser and external artefacts can address the criticisms of EM, such as cognitive bloat and the coupling-constitution fallacy.

This chapter is organised as follows. Section 2.2 focuses on the key principles of EM. Section 2.3 looks at criticisms of the EM thesis, in particular, two predominant criticisms: (1) the coupling-constitution error and (2) parity and functionalist poise. In Section 2.4, I analyse Clark's position within the broad spectrum of cognitive science and the philosophy of mind.

I conclude that C&C's original proposal for EM without functionalist poise has the potential to develop in new directions. Therefore, my aim is to modify C&C's original EM thesis to make it immune to the coupling-constitution fallacy and cognitive bloat. My conclusions are given in Section 2.5.

2.2 The EM Hypothesis

I asked my daughter another question: "If you use a pencil and paper to accomplish a complex mathematical task, where does the computation occur that produces the results of that mathematical task, i.e. the mental state or beliefs derived from that cognitive task? Are they inside your brain, internal to your physical body or do they spill over to the pen and paper in a closely coupled way with your brain and body?" Her reply was the same as before: "My mind is within me, inside my physical boundary. However, the pencil and paper helped me to solve the problem." The EM hypothesis is counter-intuitive, as it does not accord with the general perception of the mind.

In this section, I explore the background and key principles of the EM hypothesis and analyse how it differs from the traditional philosophy of mind, especially how 4E cognition, which stands for embodied, extended, enactive and embedded cognition, is a contemporary framework for understanding cognition that goes beyond the traditional view of the mind as a disembodied information processor located solely in the brain. This perspective emphasises the interaction between the mind, the body and the environment and the context in which cognition takes place.

There is some overlap between cognitive processes and mental states. The hypothesis of extended cognition (HEC) claims that cognitive processes extend beyond the skin bag of the cognitive agent. Similarly, the EM thesis asserts that mental states extend beyond the skin bag of the cognitive agent. Since some extended cognitive processes are belief-forming cognitive processes, the resultant beliefs and mental states are extended.

Semantics refers to the mental states by which we assign meaning to words, sentences and symbols. The associated cognitive processes can be (1) language processing, i.e. understanding the meaning of words and constructing meaningful sentences; (2) memory, i.e. storing and recalling the meaning of words or concepts; and (3) reasoning, i.e. applying meaning in a logical deduction or analysis. A propositional attitude refers to a mental state or stance that one holds towards a proposition and includes belief, desires, hopes, fears and intentions. The associated cognitive processes can include decision-making, namely, making decisions based on beliefs or desires about a proposition. For example, if someone believes it is going to rain (a propositional attitude), they may decide to carry an umbrella. There is a subtle difference between the extension of cognitive processes and the extension of the mind. The difference is mainly associated with the processes and the resulting state. The underlying assumption is that extending the mind requires extending cognitive processes.

The traditional concept of mind, i.e. that the mind is inside the skull and skin bag, can be traced back to the Cartesian concept of mind and body. Contemporary philosophies of mind, such as functionalism, are aligned with this concept of mind. Functionalism claims that mental states and cognitive processes are constituted by the subpersonal causal roles played by sensory inputs, intermediate mental states and behavioural outputs. A particular mental state can be realised by input and output relations. A pain state can be caused by an input due to damage to any part of the body, and the resulting output could be moaning or crying. The fact that creatures with an anatomy and physiology radically different from us, such as an octopus, can show pain behaviours, suggests that pain cannot be equated with having a particular neural or chemical property, i.e. a mental state can be realised in different ways. The same functionality can be realised in different ways. It is analogous to the example of a chair. The functional role of a chair is for sitting on. However, a chair can be made from many different types of material, such as plastics and metal. In functionalism, each mental state can be specified by its relation to other mental states and the inputs and outputs that it deals with. For example, we may have a belief that tigers are dangerous. What could cause someone to believe that tigers are dangerous? Here, the input may have been a documentary or a book that teaches us that tigers can eat us. Alternatively, we may have learned that powerful animals with big teeth are dangerous. From these mental states or beliefs as input, we then form the belief that tigers are dangerous. The behavioural outputs may be cautious behaviour. For example, if we are near a tiger, then we might act cautiously such as by moving away. We may warn others that we have spotted a tiger, or we might decide not to visit places where wild tigers roam without taking safety precautions. This mental state, i.e. the belief that tigers are dangerous, can output other mental states, such as feelings. For example, we may experience fear or anxiety when thinking about or encountering a tiger or we may form a desire to learn more about tigers to better understand them. If we see a tiger, we may have a desire to run away from it, we may actually run away from it, or we may get a weapon for defence. Here the mental states are individuated by their functional roles.

In functionalism, what makes the belief "tigers are dangerous" a specific kind of mental state is not a particular set of neurons firing in the brain when this belief is being considered, but rather, it is the way that this belief interacts with sensory inputs, produces behavioural outputs and affects other mental states. In other words, it is the role that this belief plays in the entire cognitive system. If an alien has a state that functions in a similar way with respect to its inputs, outputs and relations, even if its internal structure is totally different from that of a human, it could be said under functionalism to have a belief that is functionally equivalent to the belief that "tigers are dangerous".

Traditional functionalism is based on mental states, constituted by the causal relations of sensory inputs, behavioural outputs, and other mental states. It is consistent with internalism and is based on the computational theory of cognition (CTC), which is a computer analogy for cognition. For example, a computer has input devices (a keyboard, mouse, touchscreen or internet connection), a processor (the central processing unit or CPU) and output devices (display monitor and speakers). In analogy, human cognition has inputs (from sensory organs), a processor (the brain) and outputs (mental states and behaviours). In CTC, mental processes are computational processes. CTC explains how mental properties can arise from physical properties via symbol manipulation. Symbols represent states that have syntactic and semantic properties. Syntactic properties are physical properties, such as shape, whereas semantic properties relate to what the symbol means or represents. Similarly, computers and calculators are symbol-manipulating machines. In a computer, the physical positions of electrons and holes in a silicon chip represent binary numbers such as 0101, which can be translated into a higher-level computer language and which can ultimately be translated into a human language, such as English, with semantic content. The mind, too, contains symbols or representations, which probably utilise the hardware of the brain, such as neurons, chemicals and electrical pulses. The elements of the brain are, under one description, neurons, but if they behave in a certain way, they might also be symbols (Shapiro, 2012). Cognitive internalism and computational functionalism share some themes. Both emphasise the importance of internal processes. Computational functionalism underscores the role of internal computational processes in producing behaviour, whereas cognitive internalism maintains that cognition is primarily an internal affair.

Some new findings in cognitive science do not accord with this traditional concept of mind, since they recognise the importance of the agent's body, their actions, the causal dependence on the environment and the constitutive role of environmental structures in cognition. Cognition is embodied when mental states and cognitive processes are constituted by bodily processes. Cognition is enactive when mental states and cognitive processes are constituted by actions. Cognition is embedded when there is an essential causal dependence between mental states and processes and the environment. Finally, cognition is extended when environmental structures can partly constitute mental states and processes.

The four E's (embodied, enactive, embedded and extended cognition) are generally considered as being part of anti-Cartesian cognitive science. They have a common theme that endorses the active role of external artefacts and contradicts the Cartesian prejudice that the mind and body are distinct and that the mind is inside the body, i.e. that the boundary of the mind is demarcated by the skin and the skull. The subtleties, differences and commonalities with four E cognition can be illustrated by the example of an experienced accountant John. While performing complex accounting calculations, John uses his notepad and a pen to offload the complex details. Instead of carrying out the entire calculation inside his head using his biological memory, he stores intermediate results in columns of figures in his notepad. The entire calculation has various repeated steps, which involve the brief use of his biological memory, his perception when he scans the columns of figures, writing in his notepad, the calculation of intermediate results and so on.

According to embodied cognition, John uses his body when carrying out the calculations. He moves his head and eyes when he scans a column and he writes in the notepad using his hand. John's body is integral to this process of carrying out the complex accounting calculations. Therefore, John's cognitive processes, i.e. performing the calculations, are embodied. When performing the calculations, John's body and mind are unified and inseparable. Therefore, John's mind and body are not distinct entities, as in traditional cognitive internalism.

According to enactive cognition, perceptions and actions are unified. Perceiving is a kind of activity. Although John may not be a good example of enactive cognition, this example can indeed be seen from an enactive perspective, which emphasises the integrated nature of the cognitive processes. The accountant is not passively reading the numbers. Instead, John is actively engaging with the data, perhaps by manipulating a physical ledger, using a calculator, making notes and so forth. His perception of the numbers and the associated actions, such as writing and calculating, are deeply intertwined. When performing a complex calculation, John scans the columns of figures and arrives at an intermediate result, which he records in his notepad. His perception of the columns of figures and his action of writing in his notepad are inseparable and unified. Perception and actions are not distinct, as in traditional cognitive internalism.

According to embedded cognition, the external artefacts (notepad and pen) have a causal role in accomplishing the complex accountancy computation. John scans the columns of figures and arrives at some intermediate results, which he stores in his notepad. Therefore, cognition is embedded in the notepad and pen, as they have a causal role in cognition. In contrast, traditional cognitive internalism considers that the notepad and pen are external, and that cognition is entirely internal to the cognitive agent.

According to extended cognition, an external artefact, such as a notepad or a pen, can have a constitutive role in accomplishing the complex computation. Without a pen and a notepad, or similar tools like a spreadsheet, John could not have accomplished the complex computation. The pen and notepad are constituent parts of the accounting calculation. The functional role of the pen and notepad in cognition is the same as the role of John’s short-term biological memory. Therefore, cognition is extended. In contrast, traditional cognitive internalism claims that cognition is internal to the agent and is not extended.

To summarise:

1. John’s cognition is embodied since he uses his body. He uses his hands to write and he moves his head to perceive and so on. His body is inseparable from his mind. Mind and body are unified.

2. When scanning the columns of figures and writing in his notepad, John’s perception and his action of writing are unified and inseparable. Therefore, his cognition is enactive since perceiving is an action. Perception and action are unified.

3. His notepad and pen have a causal role in accomplishing the task; therefore, his cognition is embedded.

4. His notepad and pen have a constituent role in accomplishing the cognitive task. Therefore, his cognition is extended.

Thus, the commonality of the four E's is that they all deny that cognition is entirely internal to the cognitive agent and that the mind and body are distinct entities. Note that there are subtle differences among the four E's. Embodied cognition considers only the embodiment of cognitive processes, i.e. the unity of mind and body. Enactive cognition is about the unity of perception and action, embedded cognition is about the causal role of external artefacts and extended cognition is about the constituent role of external artefacts in cognition.

One of the important features of embodied cognition is its emphasis on the dynamic nature of the cogniser. Cognitive behaviour is the product of interactions between the brain, the body and the world. The application of DST in cognition can describe how cognitive behavioural processes emerge from the interaction of multiple components over time. Embodied cognition and DST have shared themes. Both emphasise the interconnectedness of an agent and their environment. Although embodied cognition highlights the importance of bodily interactions in shaping cognition, DST provides tools and concepts for modelling and understanding this interaction in terms of dynamic systems. A dynamical system comprises tightly coupled components such that changes to each component over time are interdependent with the state of the other components and the overall state of the system. An example of a dynamic system is the solar system. If the positions of the sun, the planets and all the other bodies in the solar system are known, then their future positions, i.e. their positions as time passes, can be predicted by Newton's laws. However, the positions of the planets are interrelated, i.e. any change to the position of one will affect the future positions of the others and vice versa. Both embodied cognition and DST challenge the traditional, computational and representational understanding of mind by suggesting that cognition emerges from an ongoing interaction rather than being a product of internal computation alone.

As mentioned above, functionalism is supported by CTC. According to computational functionalism, the belief that "tigers are dangerous" is represented in the brain as stored information that can be activated or retrieved in relevant contexts. This encoded information can interact with other beliefs and sensory data to produce behavioural outputs. Moreover, cognitive processes and mental states do not change over time. However, in a dynamic system, the relations between the components of the system can change over time. Thus, for the belief that tigers are dangerous, many aspects need to be continually considered, such as the agent’s bodily motion over time, how the agent perceives a tiger over time, visual invariants, the interactions of that mental state with other beliefs over time, the possible outcomes over time such as a desire to run and the feedback loop involved in perceiving the current position of the tiger. The belief that “tigers are dangerous” can be seen in computational functionalism as stored information that is activated in relevant contexts or in DST as an emergent property of various dynamic interactions over time. DST provides a more holistic understanding of cognition as it emphasises its adaptive and evolving nature. It accounts for the fact that beliefs and behaviours can change based on context, rather than fixed inputs leading to fixed outputs, as in computational functionalism.

Tim van Gelder is known well for his dynamical systems approach to cognition. Van Gelder (1995) proposes a notable analogy by comparing cognition to a Watt governor. The operation of a Watt governor, as explained below, provides insights about the application of DST in cognition. Although a Watt governor does not itself think or cognise, its principle of operation can provide insights into how DST conceptualises cognition. It highlights the emergent, adaptive and feedback-driven nature of cognitive processes. Now, the future state of a dynamic system depends on its current state. In DST, systems can be modelled mathematically using differential equations that describe how the parameters of the system change over time. The Watt governor on a steam engine is an example of a dynamical system (Figure 2.1).

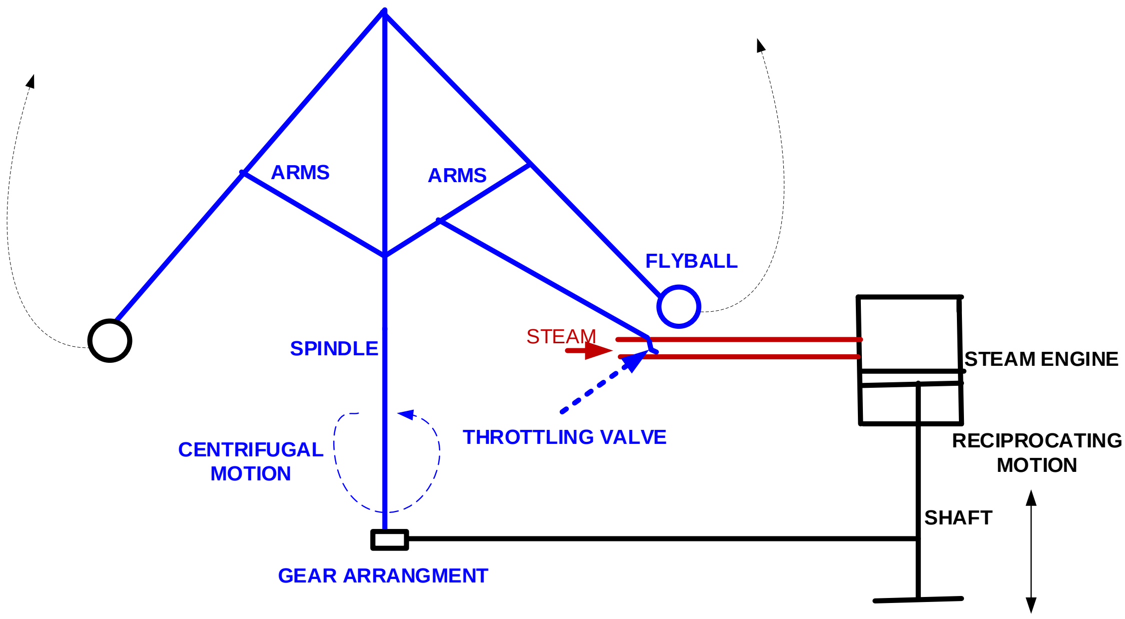


Figure 2.1. Watt governor.

The aim of the Watt governor is to keep the shaft driven by a steam engine rotating at a constant speed. This governor has two flyballs, each of which is connected to the spindle by two hinged arms. The flyball arms are connected to a throttle valve that controls the amount of steam sent to the engine. The spindle is directly connected to the rotating shaft. If the speed of the shaft increases, the spindle rotates faster and the flyballs move outward due to the centrifugal force and thus, upward because of the way the arms are connected. This partially closes the throttle valve, which reduces the amount of steam going to the engine, which, therefore, slows down. As the shaft speed reduces, the spindle rotates more slowly and so the flyballs move back in, which reopens the throttle valve. In this way, the flyball governor controls the speed of the engine. The balls, the valve and the shaft are coupled together and the state of any one of these components includes information about the states of the others. The observation that a Watt governor operates without a distinct beginning, middle or end is indicative of its continuous cyclical feedback mechanism. In DST as it relates to cognition, this continuous nature has parallels for understanding the ongoing, adaptive and emergent nature of cognitive processes. Just as a Watt governor continually adjusts itself based on the engine’s speed, DST suggests that cognitive processes are in a constant state of responding to feedback. Cognition can be considered to be a dynamic system in which the brain, body and environment are closely coupled. Cognition takes place in time and is dependent on the environment, on the mental state and on sensory feedback from the body.

CTC does not incorporate dynamism among the brain, the body and the environment. Instead, it considers that all cognitive activity occurs within the brain and depends solely on the computational processes of symbolic manipulation. Computational processes have a beginning, a middle and an end. Unlike dynamic systems, sequences such as input–output are important for computational process. In the above example of the Watt governor, structurally, we can say that there is an input end (the spindle) and an output end (the throttle valve). However, the entire operation of the governor is smooth and continuous, and there are no discrete steps and no beginning or end. The process is completely continuous and all components and actions are interrelated (Van Gelder, 1995):

> Computationalists standardly think of a process as commencing with an input to the system. The task for the system is to produce an appropriate output, and it does so via a sequence of internal operations culminating in the system's halting with that output. Dynamicists, by contrast, think of processes as always ongoing, not starting anywhere and not finishing anywhere. The goal is not to map an input at one time to an output at some later time, but to constantly maintain appropriate change. (Van Gelder, 1998, p. 7)

C&C claim that the EM thesis is supported by a growing body of research into cognitive science:

> In areas as diverse as the theory of situated cognition (Suchman, 1987), studies of real-world robotics (Beer, 1989), dynamical approaches to child development (Thelen & Smith, 1994) and research on the cognitive properties of collectives of agents (Hutchins, 1995), cognition is often taken to be continuous with processes in the environment. (C&C, 1998, p. 10)

According to this new research, an agent's cognitive processes are inseparable from the active interaction of the agent with their environment. Their cognitive processes are situated, embodied, dynamic, enactive and distributed. C&C's EM hypothesis claims that the mind is not entirely internal to the agent but at times it partly extends into the environment, such as when an external artefact plays a constitutional role in cognition. Thus, the external environment plays a constitutive role in cognition and the production of beliefs. Extended beliefs are part of the mind.

The EM thesis is built on epistemic action. Section 2.2.1 explains the difference between epistemic action and pragmatic action.

2.2.1 Epistemic and Pragmatic Actions: The Tetris Player

Kirsh and Maglio (1994) differentiate between pragmatic actions and epistemic actions. An epistemic action does not have a physical goal to achieve but gets information. In contrast, the intention of a pragmatic action is to achieve a physical goal. The demarcation between pragmatic and epistemic actions is often not clear, as some pragmatic actions can get information as well as achieving a physical goal. Someone thirsty getting a cold drink from a fridge is an example of a pragmatic action. In this action, there is no intention to gather information. However, an epistemic action is exploratory, such as looking inside the fridge to see what ingredients there are for cooking dinner, and there is no immediate practical goal. According to Kirsh and Maglio (1994), the primary function of a pragmatic action is to bring the agent closer to a physical goal. Thus, pragmatic actions alter the world because some physical change is desirable for its own sake, e.g. putting cement into a hole in a dam to fix a leak (C&C, 1998, p. 8). However, in an experiment with a Tetris player, Kirsh and Maglio (1994) identify that some zoid rotations made by the Tetris player were not pragmatic actions, i.e. they were not made to achieve the goal of fitting the zoid into the wall at the bottom of the grid. Rather, such rotations were used to gain information about the zoid, to see how the zoid could fit into the wall etc. Kirsh and Maglio argue that such actions are epistemic actions, which are physical actions that make mental computations easier or make a problem-solving task simpler. The Tetris player can solve the necessary cognitive and perceptual problems more quickly by acting in the world than by relying solely on computations performed within their head. Clark provides examples of epistemic and pragmatic actions:

> Walking to the fridge to fetch a beer is a pragmatic action. Epistemic actions may or may not yield such physical advance. Instead they are designed to extract or uncover information. Looking inside the fridge to see what ingredients are available to cook tonight's dinner is a mild species of epistemic action. (Clark, 2008, p. 71)

Kirsh and Maglio (1994) state that epistemic actions "are best understood as actions that use the world to improve cognition. These actions are not used to implement a plan, or to implement a reaction; they are used to change the world in order to simplify the problem-solving task." Epistemic actions unite action and cognition, in contrast to the traditional view that action and cognition are distinct. As we can see, Kirsh and Maglio claim only that epistemic actions aid and improve performance on certain cognitive tasks. However, C&C (1998) extend this claim by asserting that epistemic actions are partially constitutive of cognition. C&C (1998) argue that: "Epistemic action, we suggest, demands spread of epistemic credit."

C&C extend the Tetris player example using a thought experiment as follows:

Case 1: A Tetris player mentally rotates a falling zoid to see how it would fit into the wall at the bottom of the grid. We call this an internal rotation.

Case 2: A Tetris player pushes a button on the screen to visualise the rotation of the falling zoids. These computations of the rotations are non-mental since they happen outside his head. We call these external rotations.

Case 3: A Tetris player has a neural implant that can compute rotations as fast as the computer in case 2. The computations happen inside his head due to the location of the neural implant.

Case 4: A Martian playing Tetris uses only his natural cognitive equipment to perform computations as fast as those in cases 2 and 3. These computations are inside his head.

The general perception is that since the Martian uses his natural cognitive equipment to compute fast rotations inside his head, these are a kind of mental rotation. The neural implant, in contrast, is artificial but it is inside the Tetris player's head. In this case, it is unclear whether the computations can be considered to be mental. According to Clark, the functional roles in all the above cases, i.e. the fast rotations of the zoids, are similar, whether there is a mental rotation or a physical rotation or whether they are computed using a neural implant or the Martian's advanced cognitive equipment. Clark (Menary, 2010, p. 44) claims that:

> In the Martian case, we would have no hesitation in classifying the fast rotations as a species of mental rotation. With this thought experiment as a springboard, we offered a parity principle. The parity principle invites us to treat the players' use of the external rotate button, the cyberpunk implant, and the Martian native endowment as all on a cognitive par. But of course there are differences. Most strikingly, in case 2 the fast-rotate circuitry is located outside the head and the results are read in by perception, whereas in cases 3

> and 4 the circuitry is all bounded by skin and skull and the results are read off by introspection.

If physically rotating the Tetris block on the screen has the same outcome in terms of fitting the block as mentally simulating the rotation, then according to the parity principle, the action of rotating the block using the game controls can be considered a cognitive process, just like the mental simulation. The key point is not where the process occurs (inside or outside the head) but the role it plays in achieving a particular cognitive outcome.

Clark concludes that case 4 is the same as case 1. Moreover, the computational processes in case 3 are performed in the same way as those in case 4. Therefore, case 3 is also equivalent to case 1. The only difference is that the computations are distributed between the agent and the computer in case 2 instead of internalised within the agent in case 3. Kirsh and Maglio conclude that the Tetris player made some external rotations of the zoids to gather information rather than to fit the zoids pragmatically into the wall. However, Clark concludes that the external rotations of the zoids made epistemically are a constituent part of the player's cognitive processes and that the internal and external rotations are functionally equivalent. There is a parity or equivalence between internal and external rotation. This parity principle is the core principle of C&C's EM thesis. The next section describes the parity principle.

2.2.2 The Spread of Epistemic Credit: The Parity Principle

C&C (1998, p. 8) claim that:

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head!

In summary, when completing a cognitive task, if a process that occurs in the external environment *functions in the same way* as an internal cognitive process, then we would not hesitate in recognising the external process as a constituent part of the overall cognitive process. Thus, since both an internal process and an external process contribute to the completion of the cognitive task, equal consideration has to be given to both, i.e. the external process and the internal process are integral parts of the overall cognitive process. The external part of the cognitive process cannot be excluded because it is not directly performed by part of the biology of the agent. This is the parity principle.

To establish and explain the main features of EM, C&C considered an Alzheimer's patient Otto, who relies heavily on his notebook for his day-to-day activities. One day, he hears that there is an exhibition at the Museum of Modern Art (MoMA) on 53rd Street. Otto wants to go to the exhibition and he records the details in his notebook. Inga, whose biological memory is normal, relies on her memory for her day-to-day activities. She also hears about the exhibition at MoMA, and she, too, wants to go to the exhibition.

On the day of the exhibition, Otto's notebook tells him that it is at MoMA on 53rd Street and he walks there. In this case, C&C argue that his memory is functionally located inside his external notebook. Without the notebook, Otto will not have any memory of the exhibition on 53rd Street. The functional roles of Inga's biological memory and Otto's notebook are equivalent. Therefore, C&C argue that Otto's memory is functionally located in his notebook. On the other hand, Inga recalls from her memory that MoMA is on 53rd Street. Comparing both people, C&C claim that the notebook has a constitutive role in Otto's cognitive processes. It has a similar functional role as Inga's biological memory. Some beliefs are occurrent beliefs, such as my belief that it is raining while I am typing this chapter. This belief arose because of the occurrence of rain outside. However, my belief that the UK's capital is London is a non-occurrent, dispositional belief. That is, this belief was already in my mind and did not arise due to the occurrence of an event. Otto has a dispositional belief about the location of MoMA. This belief cannot be stored in his biological memory due to his Alzheimer's. Instead, he relies on a notebook. When he needs this information, he has a disposition or tendency to consult his notebook. Thus, Otto's dispositional belief is functionally similar to Inga's dispositional memory about MoMA. C&C argue that some mental states, like experience, are determined internally, but that some beliefs are constituted by features of the environment, such as Otto's dispositional belief based on his notebook that the museum is on 53rd Street. Therefore, in this case, his mind extends into his environment. Since Otto's cognitive processes use the external memory in his notebook, which has a constitutive role, C&C claim that Otto's cognition extends into his environment. Section 2.2.2.1 summarises the key arguments for the EM thesis.

2.2.2.1 Arguments for EM

P1. In some cognitive processes, external artefacts are coupled with the cognitive agent. "The human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right" (C&C, 1998, p. 8). An example of coupling is Otto and his notebook.

P2. The coupling with an external artefact can lead to its having an active role in cognitive processes, if the coupling of the external artefact with the cognitive agent meets the following conditions: "(a) All the components in the system have an active causal role. (b) They jointly govern behaviour in the same way that cognition usually does. (c) If we remove the external component, then the system's behavioural competence will drop, just as it would if we removed part of someone's brain" (C&C, 1998, p. 6). This sort of coupled process is equivalent to a cognitive process, whether or not it is wholly in the head (C&C, 1998). Otto's notebook plays an active role in Otto's belief-forming processes.

P3. The coupling of an external artefact with the agent can lead to its having a constitutive role in cognitive processes, if the coupling meets the following conditions: (a) The resource must be reliably available and typically invoked. (Otto always carries his notebook and will not answer that he "doesn't know" until after he has consulted it.) (b) Any information thus retrieved must be automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed as trustworthy as something retrieved clearly from one's biological memory. (c) The information within the resource must be easily accessible as and when required. (d) The information in the notebook has been consciously endorsed at some point in the past.

P4. If an external artefact has a constitutive role in cognition and is functionally equivalent to its biological counterpart, there is no hesitation in asserting that the external artefact has a role in cognitive processes. For example, Otto's notebook serves as his memory regarding the location of MoMA on 53rd Street and is functionally equivalent to Inga's biological memory. Otto's notebook and Inga's biological memory are functionally equivalent.

> If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part

> of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head.
>
> (C&C, 1998, p. 29)

Thus, Otto's cognition is extended.

P5. Beliefs are part of the mind. Some beliefs are formed by extended cognitive processes. Otto's cognition extends into his notebook, so his non-occurrent dispositional beliefs about MoMA on 53rd Street also extend into the notebook.

C1. Since Otto's belief is extended, Otto's mind is extended.

In Section 2.2.3, I am going to discuss the cornerstone for the EM thesis, namely active externalism, and explain how it differs from content externalism.

2.2.3 Active Externalism

The debate between cognitive externalism and internalism in the philosophy of mind is about the location of cognitive processes and mental states. Internalists argue that cognitive processes and mental states are wholly internal to the agent, whereas externalists deny this claim. In extended cognition, externalists claim that cognitive processes can extend into the environment if an external artefact is a constituent part of cognitive processes. Therefore, cognition is not always entirely internal, as the internalists claim.

C&C (1998, p. 1) ask: "Where does the mind stop and the rest of the world begin?" This has two standard replies: (1) The demarcation of the skin and skull. Thus, what is outside the body is outside the mind. (2) It "just ain't (all) in the head". They hold that this externalist meaning carries over into an externalism of the mind. C&C (1998, p. 4) propose a third position, "a very different sort of externalism: an active externalism, based on the active role of the environment in driving cognitive processes".

C&C propose a new form of externalism, active externalism, to substantiate their claim that the mind can extend into the environment. Active externalism considers the immediate role of the environment in the production of beliefs and behaviours, as in the case of the Tetris player and Otto. According to C&C, the augmented external constituent of a cognitive process demonstrates active externalism, i.e. real-time externalism, and there is no need to consider history and location. According to active externalism, the external environment can play an active role in cognitive processes, as with Otto's notebook and the Tetris player rotating the zoids. According to HEC, the environment plays a constitutive role in cognitive processes, such as in the external manual rotation of zoids by the Tetris player. The objective of the internal mental rotation of zoids is to collect information about the shape, type and characteristics of the contours they could fit into. A physical external rotation of a zoid by a Tetris player is also to gather information about where it can fit. Therefore, an external physical rotation is functionally equivalent to an internal mental rotation, since both unveil more information about the zoids. The production of behaviour and beliefs depends heavily on the role of the external artefact. Therefore, HEC implies active externalism. The Tetris player demonstrates the extension of cognition, i.e. cognitive processes extend into the environment when environmental structures partly constitute cognitive processes. Moreover, Otto demonstrates the extension of mental states, i.e. dispositional beliefs extend into Otto’s notebook.

Clark (2008b) states that EM is a claim “about extended vehicles – vehicles that may be distributed across brain, body and world”. Therefore, EM is a claim about the physical processes that realise mental states and cognitive processes.

Hurley (2010) differentiates between what-externalism, which appeals to external factors to explain the what of mental states, and how-externalism, which appeals to external factors to explain the how of mental states. In Hurley’s taxonomy, traditional content externalism is what-content externalism whereas the EM hypothesis, which is based on active externalism and causal coupling, is how-content externalism or content-enabling externalism. According to Hurley, externalism is explanatory rather than ontic.

In Section 2.2.4, I discuss an important aspect of the EM thesis: the causal coupling between an artefact and a cognitive agent.

2.2.4 Causal Coupling

C&C claim that external coupling is part of our core cognitive resources because “the biological brain has in fact evolved and matured in ways which factor in the reliable presence of a manipulatable external environment” (C&C, 1998, p. 11). Evolution favoured offloading some of our cognitive load to the environment, and as a result, the environment is part of the cognitive loop. C&C state that language is such a coupled system in which cognitive processes extend into the world. Without language, we would have Cartesian inner minds in which high-level cognition would largely depend on our internal resources.

C&C (1998) argue that:

> The human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right. All the components in the system play an active causal role. They jointly govern behaviour in the same sort of way that cognition usually does. If we remove the external component, then the system's behavioural competence will drop, just as it would if we removed part of someone's brain.
>
> C&C (1998, p. 8)

C&C assert that this sort of coupled process is equivalent to a cognitive process, whether or not it is wholly in the head. As an example, Otto's notebook is closely coupled to his behaviour. If Otto's notebook is not available to Otto, his beliefs about the exhibition at 53rd Street will be impaired, and he will not go to the exhibition. Otto and his notebook are so closely coupled that the coupling governs Otto's behaviour. The memory stored in Otto's notebook plays the same role as Inga's biological memory. Therefore, in the cognitive processes, the functional role of memory in the notebook is equivalent to the functional role of Inga's biological memory.

In Section 2.2.5, I will discuss the conditions (namely, the trust and glue conditions) necessary for an artefact to be part of extended cognitive processes.

2.2.5 Trust and Glue Conditions

There are some notable differences between the cognitive process associated with the brain and those with external artefacts. The brain is reliable and always integrated with the agent. However, external artefacts, such as pen, paper or a smartphone, can easily be decoupled from the agent and may not be available when required. Therefore, an external artefact is decouplable and portable and not as reliable as the brain. Moreover, C&C (1998) argue that if there were no limiting conditions for the constitutive role of an external artefact in cognition, any external artefact could be considered as having a part in cognitive processes, such as an entire telephone directory or the internet. Counter-intuitively, such an unlimited extension of external artefacts into cognitive processes would mean that cognitive processes can extend anywhere via external artefacts without any limits. Thus, there is the potential for the rampant expansion of cognition into the environment. For example, does my cognitive state somehow spread across the internet? C&C consider that certain conditions are needed for an external artefact, such as Otto's notebook, to count as a constituent part of a cognitive process. These conditions circumscribe the nature of belief and the way that we incorporate an artefact into our cognitive processes. The conditions are not arbitrary. As C&C (1998) argue, they help us to understand what is involved in the ascription of an extended belief. These conditions are somehow related to common-sense functionalism, which is the idea that mental states are defined by their functional role in a system rather than by their intrinsic properties. C&C's extended mind thesis can be seen as a kind of functionalism because it argues that if an external object or process plays the same functional role as a cognitive process inside the brain, then it should be considered a part of the mind. The conditions that C&C set out can be seen as specifying the functional roles required for something to be considered a belief.

According to C&C (1998, p. 17), the following conditions must be satisfied:

1. The notebook must be a constant in Otto's life. The information in it is vital and he will rarely act without consulting it.

2. The information in the notebook is always directly and easily available.

3. Upon retrieving information from the notebook, he endorses it automatically.

4. The information in the notebook has been consciously endorsed, and indeed, it is there as a consequence of this endorsement.

Clark (2008, p. 79) introduces conditions that an external artefact must satisfy if it is to be considered a constituent part of a cognitive process. These are known as the trust and glue conditions:

1. The resource must be reliably available and typically invoked. (Otto always carries his notebook and will not answer that he "doesn't know" until after he has consulted it.)

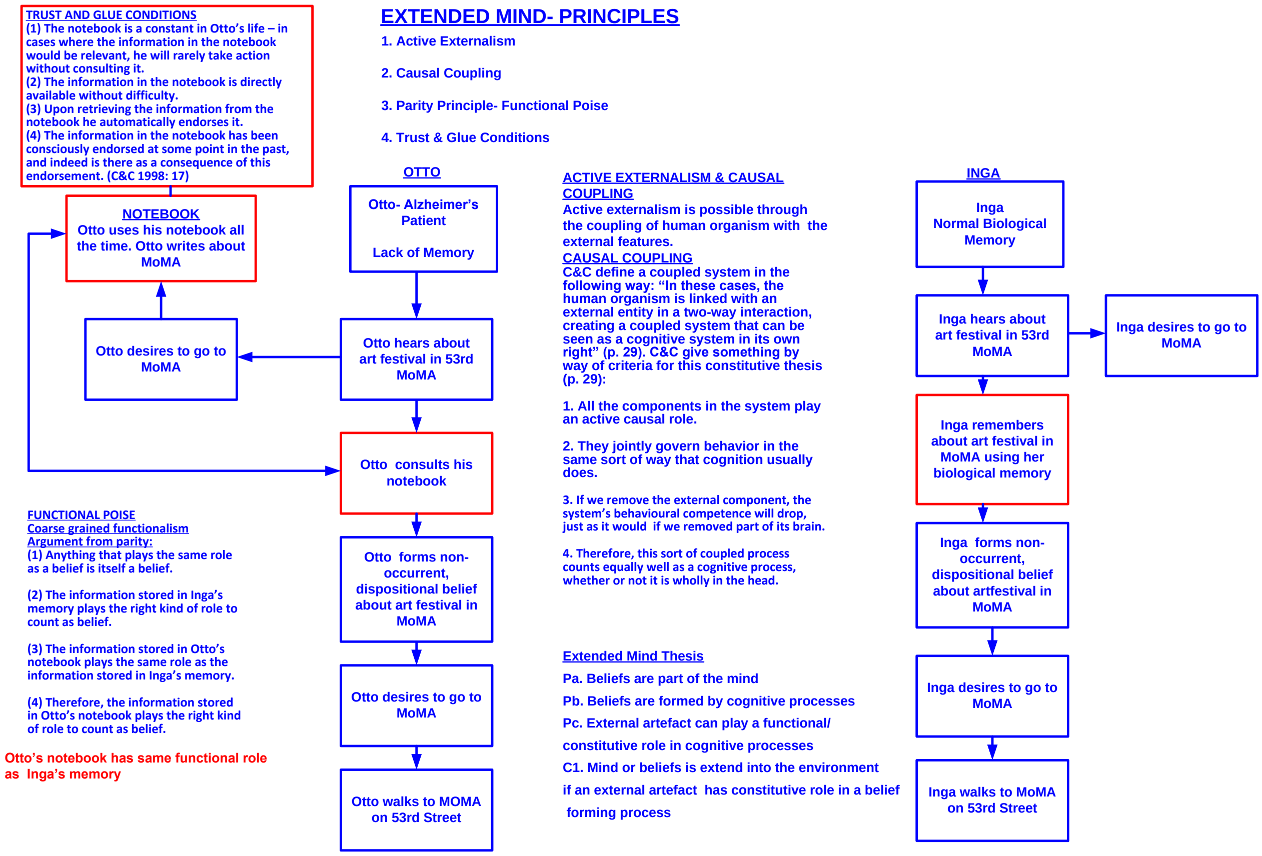
2. Any information thus retrieved must be automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed as trustworthy as something retrieved clearly from one's biological memory.

3. The information within the resource must be easily accessible as and when required.

4. That information in the notebook has been consciously endorsed at some point in the past.

The fourth condition means that for an external artefact (e.g. Otto’s notebook) to be considered a genuine part of one’s extended cognitive system, the information contained within it must have been consciously accepted or approved by the user at some point in the past. This fourth condition emphasises the active involvement of the individual in integrating external tools or resources into their cognitive processes, just like Otto must have made a conscious decision to trust and record specific pieces of information in his notebook. The conscious endorsement condition aligns with the parity principle, as ensuring that the external resource is consciously endorsed should consequently ensure that the artefact is not just any arbitrary tool but an integral part of the cognitive process, making it easier to see it as functionally equivalent to an internal process. However, this functional equivalence can be challenged, as the conscious endorsement adds an additional layer of requirement that is not required for internal cognitive processes because not all the beliefs that are stored in one’s brain are consciously endorsed or scrutinised. Some beliefs may be implicitly learned or absorbed without conscious reflection. From this perspective, the fourth condition is not entirely consistent with the parity principle.

Figure 2.2 shows the principles of EM. The red boxes on the left show Otto, Otto’s notebook, and the trust and glue conditions, which together serve as Otto’s memory. They are equivalent to the red box on the right, which represents Inga’s biological memory. Since Otto’s notebook is an external artefact and plays a key role in Otto’s beliefs, just like Inga’s beliefs based on her biological memory, C&C conclude that Otto’s beliefs extend into his environment, as does his mind.



TRUST AND GLUE CONDITIONS
(1) The notebook is a constant in Otto's life – in cases where the information in the notebook would be relevant, he will rarely take action without consulting it.
(2) The information in the notebook is directly available without difficulty.
(3) Upon retrieving the information from the notebook he automatically endorses it.
(4) The information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement. (C&C 1998: 17)

EXTENDED MIND- PRINCIPLES
1. Active Externalism
2. Causal Coupling
3. Parity Principle- Functional Poise
4. Trust & Glue Conditions

ACTIVE EXTERNALISM & CAUSAL COUPLING
Active externalism is possible through the coupling of human organism with the external features.
CAUSAL COUPLING
C&C define a coupled system in the following way: "In these cases, the human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right" (p. 29). C&C give something by way of criteria for this constitutive thesis (p. 29):

1. All the components in the system play an active causal role.

2. They jointly govern behavior in the same sort of way that cognition usually does.

3. If we remove the external component, the system's behavioural competence will drop, just as it would if we removed part of its brain.

4. Therefore, this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head.

FUNCTIONAL POISE
Coarse grained functionalism
Argument from parity:
(1) Anything that plays the same role as a belief is itself a belief.

(2) The information stored in Inga's memory plays the right kind of role to count as belief.

(3) The information stored in Otto's notebook plays the same role as the information stored in Inga's memory.

(4) Therefore, the information stored in Otto's notebook plays the right kind of role to count as belief.

Otto's notebook has same functional role as Inga's memory

Extended Mind Thesis
Pa. Beliefs are part of the mind
Pb. Beliefs are formed by cognitive processes
Pc. External artefact can play a functional/ constitutive role in cognitive processes
C1. Mind or beliefs is extend into the environment if an external artefact has constitutive role in a belief forming process

NOTEBOOK
Otto uses his notebook all the time. Otto writes about MoMA
Otto desires to go to MoMA
OTTO
Otto- Alzheimer's Patient
Lack of Memory
Otto hears about art festival in 53rd MoMA
Otto consults his notebook
Otto forms non-occurrent, dispositional belief about art festival in MoMA
Otto desires to go to MoMA
Otto walks to MOMA on 53rd Street
INGA
Inga Normal Biological Memory
Inga hears about art festival in 53rd MoMA
Inga desires to go to MoMA
Inga remembers about art festival in MoMA using her biological memory
Inga forms non-occurrent, dispositional belief about artfestival in MoMA
Inga desires to go to MoMA
Inga walks to MoMA on 53rd Street

Figure 2.2. Extended mind principles

2.3 Criticisms of EM

This section discusses some of the criticisms of the EM thesis. The strategy I use is to put the arguments for the EM thesis first and then describe criticisms of those arguments. The arguments for EM can be summarised as follows:

P1. The external environment plays an active causal role in some cognitive processes in accomplishing a cognitive task. The human organism can be linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right (causal coupling).

P2. In the completion of a cognitive task, if an external artefact has the same functional role as an internal cognitive process, then it can be recognised as being part of the overall cognitive process (parity principle).

P3. An external resource can be considered as a constituent of cognition only if the resource is reliably available and typically invoked. Any information thus retrieved must be automatically endorsed, and the information within the resource must be easily accessible, as and when required (trust and glue conditions). An external artefact that meets the trust and glue conditions for cognitive processes has a constituent role in the accomplishment of a cognitive task. Therefore, cognition extends beyond the skin bag and skull. Note that these conditions are necessary but not sufficient. They may not be able to capture all the nuances or complexities of our relationship with external tools and resources. Even if a particular artefact meets the criteria, not all such artefacts can be considered as an extension of our mind. There are varying degrees to which tools and resources are integrated into our cognitive routines. An artefact that is deeply integrated, like Otto's notebook, could be different from a tool that is used occasionally, even if both are automatically endorsed and easily accessible.

P4. Beliefs are part of the mind. Some beliefs are formed by extended cognitive processes.

C1. Therefore, based on the above premises, it follows that the mind can extend beyond the boundaries of the individual's body and into the world by incorporating external resources as integral components of cognitive processes.

Since the inception of the EM thesis, several criticisms have been raised against it, mainly focusing on the distinction between the biological nature of cognitive processes versus external cognitive resources, the causal role of external cognitive resources versus the constitutive role of external cognitive resources, and the portability and reliability of the brain/body versus those of unreliable external cognitive resources.

In the next section, I summarise various criticisms of the EM thesis. In addition, I consider in detail two predominant criticisms of the EM thesis: (1) the coupling-constitution error and (2) parity and functionalist poise.

2.3.1 The Coupling-Constitution Fallacy

This section evaluates the criticisms against P1 and P3, which relate to the nature and constitution of the coupling of external artefacts in cognition.

The externalist claim for the constitutive role of external artefacts in cognition has been challenged by internalists. There are issues with the boundary and nature of cognition, such as for non-derived content and for laws associated with psychological states. Non-derived content does not require the independent or prior existence of other content, since it is naturalistically determined. However, derived content is conventionally determined. By defining the characteristics and nature of cognition, A&A and Rupert note that intentional content, scientific laws and the laws of cognitive processes are incompatible with the fleeting processes that rely on external artefacts.

According to A&A (2001, 2010) and Rupert (2004, 2010), there are three issues with P1 and P3:

1. The enabling or causal role of an external artefact is sufficient to explain cognition, and a constitutive role is unnecessary. A&A assert that this is a *coupling-constitution error* made by EM theorists.
2. A&A argue that if an external artefact is a constitutive part of a cognitive process, then the external artefact has to be cognitive. They note that only the *mark of cognition* can be used to determine what is part of a cognitive process. Thus, A&A suggest that non-derived content is the mark of cognition and argue that non-derived content is internal to the cognitive agent.
3. As with P3, if an external artefact can be considered as part of cognition because it has a constitutive role, then many external artefacts could be considered as being part of cognition, resulting in *cognitive bloat*.

An external artefact, such as a pen or paper, can be an enabler of cognitive processes. These artefacts can help to accomplish the objective of a cognitive process, such as a complex computation. In this case, what is the role of the pen and paper in the complex computation? Are the pen and paper a constitutive part of cognition? Or do the pen and paper merely enable the complex computation? For an EM theorist, the pen and paper are coupled to the mathematician; therefore, the pen and paper are constituent parts of the cognitive process. However, A&A argue that the pen and paper have only an enabling or helping role in the accomplishment of a complex mathematical computation. The pen and paper cannot be considered as part of cognitive processes since cognitive processes are wholly biological, i.e. the agent's brain and body are the sole owner of their cognitive processes. A non-biological external artefact, such as a pen or paper, cannot be considered a constituent part of cognitive processes. A&A claim that C&C are wrong by considering that the enabling role of pen and paper has a constitutive role in cognitive processes. This error is called the coupling-constitution error. A&A argue that EM theorists are making a coupling-constitution error when they suggest that the causal or dependent role of external artefacts is a constitutional role. If an external artefact Y is causally dependent when accomplishing cognition process X, that does not mean that cognition X is part of Y, the external artefact. A&A (2010) assert that even an acceptance of the coupling of external features of cognition does not mean that cognition extends to every part of that system.

Rupert (2004) raises concerns about the constitutive role of external artefacts. Rupert admits that external artefacts, such as a pen and paper, have an enabling role when performing a complex computation but that does not mean that the pen and paper are part of the overall cognitive process. Rupert argues that cognitive processes are wholly internal to the agent They are biological and occur in the brain and body, so that no external artefact can be considered as being part of cognition. Rupert argues that extended cognition is not consistent with traditional cognitive science, which asserts that cognitive processes are internal to the agent. Rupert's arguments highlight the challenges in forming scientific laws and in generalising about the role of external artefacts in cognition. If cognition is to be understood in terms of systematic and law-like behaviours and if external devices are inherently unreliable and decouplable, then it becomes challenging to incorporate them into a unified, scientific understanding of cognition. According to Rupert, if something is to be genuinely considered cognitive, then it should exhibit a certain degree of reliability and permanence. The fleeting nature of interactions with an external artefact and the susceptibility of an artefact to various forms of failure make artefacts unsuitable for what is traditionally considered to be cognitive. It is difficult to make scientific laws about the role of external artefacts in cognition because these external artefacts are decouplable and unreliable. To emphasise the importance of external artefacts in cognition, Rupert (2010) proposes the hypothesis of embedded cognition (HEMC) in contrast to the EM thesis. Rupert (2004) argues that embedded cognition offers a better explanation than EM for the role of external features in cognition. In embedded cognition, cognitive processes depend very heavily on organismically external properties and devices and on the structure of the external environment in which cognition takes place. HEMC endorses the causal role of and dependence on external artefacts, unlike the constitutive role.

Clark (2008, p. 130) notes:

> The error of objecting that externalist explanations give a constitutive role to external factors that are "merely causal" while assuming without independent argument or criteria that the causal/constitutive distinction coincides with some external/internal boundary. To avoid thus begging the question, we

should not operate with prior assumptions about where to place the causal/constitutive boundary, but wait on the results of explanation.

2.3.1.1 Mark of Cognition

C&C (1998) define cognitive processes through the parity principle. However, critics like Rupert and A&A do not accept this definition of cognitive processes, which they consider to be wholly biological and internal to the agent. Rupert argues that the biological part and the extended part of extended cognitive processes are different and both cannot properly be called cognitive (Rupert, 2004). The criticism of EM and C&C's responses show that there are differences in the definitions of what can be considered to be a cognitive process. I argue that these differences do not challenge EM, since EM does not claim that Inga's biological memory and Otto's notebook (i.e. external memory) are the same. Instead, it considers that the constitutive part of external memory has a part in the whole process of cognition (Clark, 2008b). EM does not claim that all cognitive processes are external. It is a hybrid approach, since it combines internal and external elements for some mental states.

All these criticisms of EM are based on internalist views, which are inconsistent with the externalist requirement for the constitution of an external artefact in cognition. For example, A&A disagree with the nature of and criteria for cognition. By relying on a causal instead of a constitutive role for his external memory in Otto's cognition, proponents of EM are forced to define what the criteria should be for a process to be considered as cognitive. As we have seen earlier, C&C (1998) provide necessary conditions. However, in many cases, these conditions are not sufficient to establish that an artefact plays a role in cognition because of the transient use of artefacts that can meet those criteria. Another issue is with the degree to which an external tool is integrated into our cognitive routines. Without a clear criterion for integration, many external artefacts could be considered as cognitive. Therefore, C&C's EM hypothesis does not conclusively state what the criteria for a cognitive process should be when establishing the constitutive role of external features.

A&A (2005) propose that non-derived content is the mark of a mental process, which they thought could be produced only by internal, biological cognitive processes. The creation of non-derived content does not require the independent or prior existence of other content, representations or internal agents. A&A propose that the content of cognitive processes requires non-derived content. They argue that an external artefact can create only derived content. A&A argue that for something to be genuinely cognitive, it must be capable of producing non-derived content. External artefacts, like notebooks and calculators, always deliver derived content since their functions and the meaning of the content depend on human interpretation. Therefore, an external artefact cannot be considered as taking part in a cognitive process. Shapiro (2009) argues that there is no conclusive theory to show how original non-derived content comes to be in the first place. Therefore, the possibility of a theory in which non-derived content is consistent with extended cognition cannot be ruled out. Ross and Ladyman (2010) note the irrelevant nature of the enquiry into the mark of cognition, since established science, like physics, does not have such boundaries. Nobody has tried to define what physics is. Menary (2010) proposes a hybrid internal and external integrationist approach for cognition that does not have a mark of cognition. Gallagher (2013) argues that the question about the mark of a mental process is irrelevant since the current debate is within the framework of the orthodox representational or functionalist model.

Establishing criteria as a mark of cognition and interpreting whether mental states meet those criteria do not support open unprejudiced research. The criteria may not be able to establish the constitutive or causal role of external features in cognition and mental states. Clark (2008b), however, agrees that being able to create non-derived content is a criterion of a cognitive process. Although Clark acknowledges that non-derived content may be a feature of certain cognitive processes, he does not claim that non-derived content is exclusive to cognitive processes, as one could still make the case that an external artefact is a part of our cognitive processes. Clark looked at a cognitive task involving a Venn diagram with intersecting sets. He argues that, irrespective of the visualisation of the Venn diagram, the meaning behind it is a matter of convention. The content derived from the visualisation can play a role in the completion of the cognitive task:

> Suppose we are busy (as part of some problem-solving routine) imagining a set of Venn diagrams/Euler circles in our mind's eye. Surely the set-theoretic meaning of the overlaps between, say, two intersecting Euler circles is a matter of convention. Yet this image can clearly feature as part of a genuinely cognitive process. (Clark, 2008, p. 48)

Clark is sceptical of A&A's claim that the content of mental processing must be composed of non-derived content. The derived content, such as the content in a Venn diagram, can be used alongside non-derived content to complete a cognitive task.

However, A&A disagree. A&A argue that visualising an intersecting Venn diagram is one thing but that understanding the meaning of a Venn diagram is another, since a Venn diagram represents existing or historical facts:

> Intersecting Euler circles on paper getting their meaning is one thing; intersecting Euler circles in mental images getting their meaning is another. Clark apparently overlooks this difference, and hence does not bother to provide a reason to think that Euler circles in mental images get their meaning via social convention. (A&A, 2010, p. 71)

A cognitive process using a Venn diagram involves three things: (1) a cogniser who can understand the patterns in the Venn diagram, (2) the pattern embedded in the Venn diagram and (3) the Venn diagram. If the cogniser cannot understand the pattern, then the visualisation of the Venn diagram will not be useful in the cognitive process. According to A&A, the meaning of the pattern within a Venn diagram is a historical fact but they do not provide any evidence for how that historical fact first emerged. A cogniser can identify a new pattern or they can identify an old pattern from a historical fact if they understand the pattern. Therefore, my argument is that an identified pattern can either have derived content, if it is historical, as in the social convention for the meaning of Venn diagrams, or underived content, if it is a new pattern, as when John Venn first introduced the concept of Venn diagrams in 1880. As Menary argues:

> If Adams and Aizawa restrict cognitive content to naturally determined contents and not conventionally determined ones then cognitive explanations will lose much of their explanatory power. If my concept of a stop sign does not contain any conventional content, then how will I know when to stop?
>
> (Menary, 2010, p. 17).

I argue that the explanatory power to account for the historical progress of the patterns embedded in external information-bearing structures should be considered as a criterion in establishing the role of external features in cognition and mental states, such as the pattern embedded in a Venn diagram. Since EM takes an integrated approach to the internal and external aspects of cognition, EM can better explain the historical progress in the patterns in external information-bearing structures and its impact on the development of tools, artefacts and social institutions.

The following sketch shows the issues raised by A&A. According to A&A, the traditional way of understanding cognition is internalist in nature. For a process to be cognitive, it must meet the mark of cognition, for example, by creating non-derived content. However, as can be seen in the sketch, neurons are physical, just like external artefacts are. It is difficult to establish at what stage of a cognitive process, from the physical neurons to the interactions with external artefacts, a mental state is produced. Somewhere in the hierarchy from neurons to the person, a cognitive system emerges that comprises any number of non-cognitive systems. At some point, a group of non-cognitive systems becomes cognitive on the next level up.

Therefore, the coupling-constitution fallacy raised by A&A against the constitutional role of external artefacts is also applicable to the subpersonal neuronal level, as there is no explanation for how physical neurons can be cognitive. A&A do not provide a response to this. Hurley (2010) also raises this concern via her magical membrane problem, i.e. how can mental states be produced in a physical brain? In Figure 2.3, the outer red box is the envelope of externalism and the purple box is the envelope of internalism.

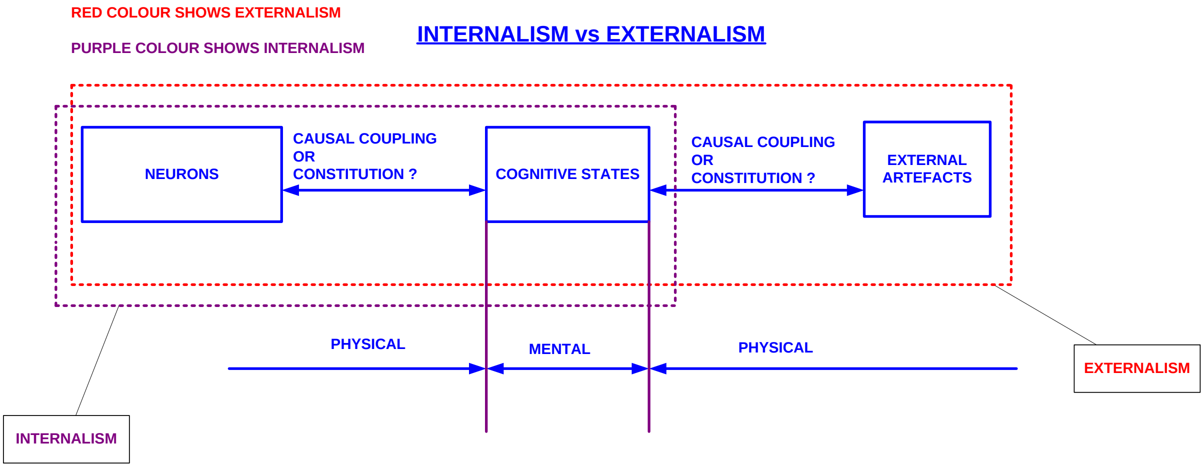


Figure 2.3. Boundaries of internalism and externalism.

According to my assessment, the explanatory power to account for the historical progress of patterns, the abstraction of thought processes, and its impact on the development of tools, artefacts and social institutions and in cognition should be considered as one of the criteria when establishing the role of external features in cognition and mental states.

As can be seen from the above, Hurley was a staunch supporter of the EM thesis. She concludes that using non-derived content as criteria for cognition begs the question of how physical neurons produce cognitive processes. According to Hurley, externalism is explanatory rather than ontic. She argues that the metaphysical argument for the constitution or composition of cognition does not help to develop cognitive science, which instead requires good explanations.

2.3.2 Cognitive Bloat

This section evaluates the criticisms against P3, which relate to the trust and glue conditions and the potential for the unlimited expansion of cognitive processes into the environment.

The claim that the mind can extend beyond the brain and body requires that external artefacts have a constitutive role in cognition. A causal or enabling role would not be sufficient to establish the extension of the mind into the world. However, even if a constitutive role of external artefacts and the extension of the mind can be established, there is a problem in demarcating the limit of that extension of the mind into the environment via external artefacts. If the words in Otto's notebook have a constitutive role in cognitive processes, then could the numbers in a telephone directory or a Google search also be a part of a cognitive process? If this is the case, then cognition would become rampantly extended via external artefacts, leading to cognitive bloat.

A&A raise concerns about P3, i.e. against the tailor-made trust and glue conditions. They argue that cognitive bloat will occur since too many external artefacts may be involved in cognitive processes. A&A (2010) raise the problem of cognitive bloat, i.e. if we accept the extension of the mind into the environment, then what is the extent of that extension? Can any external object be considered as part of cognition? C&C gave criteria to distinguish between extended cognitive and external non-cognitive processes. The resource behind an extended cognitive process should be reliable and accessible. There should be automatic endorsement of the information retrieved and that information should previously have been consciously endorsed by the subject. Clark (2008b, p. 80) argues that:

> Applying the four criteria yielded a modestly intuitive set of results for putative individual cognitive extensions. A book in my home library would not count. The cyberpunk implant would. Mobile access to Google would not (it would fail condition 2). Otto's notebook would. Other people typically would not (but could in rare cases) – and so on.

There are three conditions that an external artefact must satisfy for it to count as a constituent part of a cognitive process: (1) The external artefact is constantly with the agent. (2) The artefact is directly available to the agent. (3) The agent automatically endorses the information from the external artefact. However, Rupert (2004, pp. 401–405) argues that the three conditions are too easy to meet. As an example, Otto may always have a telephone directory with him. It is directly available to him and automatically endorsed by him. Since the three conditions have been met, then it could be argued that Otto's cognitive processes extend into the entire telephone directory, which is counter-intuitive. To avoid such a rampant expansion of cognitive processes, C&C added a fourth condition: the information in the external artefact must have been consciously endorsed by the agent at some point in the past. Since Otto has not consciously endorsed all the numbers in the directory, there is no rampant expansion of the cognitive processes into the entire directory. In response, Rupert challenges the fourth condition:

> Past-endorsement criterion undermines what is supposed to be, if one accepts HEC theorists' revolutionary sounding rhetoric, one of the most important theoretical implications of HEC: that there is no good reason to assign special status to the boundary between organism and environment. If an extended (or any) belief requires conscious endorsement in order to be a genuinely held belief, and conscious endorsement is ultimately an internal process (that is, one that takes place within the organismic boundary), then the traditional subject is privileged in a deep sense, after all. (Rupert, 2004 p. 404)

Since the conscious endorsement of information is internal to the agent then the extension of cognitive processes based on such past endorsement will not establish that the cognition is not limited to organism's boundary. As we have seen earlier, the past-endorsement criterion is only one of the criteria that Clark formulated for cognitive extension, as the others require accessibility, reliability, availability and automatic endorsement. These criteria are applied together to determine whether an external artefact is genuinely part of the extended cognitive system. According to Clark, if an external artefact, like Otto's notebook, plays a recurring functional role in Otto's life and if Otto has a history of integrating and endorsing that notebook, then the notebook is part of his extended cognitive system.

Rowlands (2009, 2010a) argues that the trust and glue conditions tailored by C&C are not sufficient to avoid cognitive bloat. By strictly following the trust and glue conditions, one can establish that any external artefact could be part of a cognitive process, which has the rampant potential to encompass every external artefact.

2.3.3 Parity, Functional Poise and the Location of the Mind

This section evaluates the criticisms against P2, which relates to the equivalence of the functional role of external artefacts and internal cognitive processes and the requirement of the parity principle. Both supporters and opponents of the EM thesis are sceptical about the parity principle and the functionalist poise required for the EM thesis:

1. Internalists disagree with the parity principle because they consider that there is a difference between internal cognitive processes and external artefacts.
2. Supporters of the EM thesis are sceptical about the parity principle as they consider that parity and functionalism are not required for the EM thesis.

Clark (2008, p. 88) claims that the EM thesis has a double appeal to functionalism: (1) it is an appeal to the common-sense or coarse-grained role implicitly grasped by normal human agents and (2) it seeks a much more fine-grained description of the actual flow of processing and representation in the (possibly extended) physical array that realises the coarse functional role itself.

The coarse- and fine-grained analysis is about the level of functional similarity required for a functional role. Since the EM thesis is based on coarse-grained similarity, there is no need for internal cognitive processes, like neural computation in the brain, and external processes, like writing in a notebook, to be similar in terms of their mechanisms or substrates. Although their mechanisms may differ, both internal and external processes can be seen as part of a cognitive system due to the similarity of their functional roles. For a fine-grained functional role, the internal and external elements need to have similar functional roles. For example, according to EM, the functional roles of biological memory and an external artefact, such as Otto's notebook, are functionally equivalent. Thus, their functional roles have a coarse-grained similarity and only equivalence is required. However, if EM were based on fine-grained similarity, then the biological memory and Otto's external memory in notebook would *need to be* similar. Note that there is always some level of similarity between Inga's biological memory and Otto's notebook since both serve a functional purpose related to cognition. On one hand, we can see a fine-grained similarity such that the mechanisms, dynamics and detailed operations need to match. On the other hand, we can see a broad, coarse-grained similarity where only the overarching function needs to align. This perspective recognises that similarity is a matter of degree, and that broad coarse-grained similarity is needed for extended cognitive processes.

Since the extended functionalist approach of EM is based on coarse-grained similarity, an external artefact can play different roles and have different properties compared to an internal biological counterpart, as in the case of Otto's notebook and biological memory. Clark (2008, p. 88) argues that the functionalist poise required for the EM thesis is a very coarse-grained common-sense functionalism and it should be differentiated from empirical functionalism: "Note that Clark's argument concerned only a subset of the folk-identified mental states, since all it requires is a form of common-sense functionalism concerning nonconscious, dispositional states. As such, the argument does not commit us to any sort of functionalism about conscious mental states" (Braddon-Mitchell & Jackson, 2007, chap. 5).

Here I will analyse the concerns of both proponents and opponents regarding the parity principle and the functionalist poise of the EM thesis. Before proceeding with this, it may be useful to see how the requirement for the spread of epistemic credit and functional poise, as detailed in C&C's 1998 paper, developed into the fully fledged parity principle and extended functionalism (2010).[1] As can be seen from Section 2.2, the requirement of coarse-grained functionalism was not explicit in C&C's 1998 paper, although Otto's case contains a clear functionalist strain. Rupert (2004) identifies this functional strain in the spread of epistemic credit and argues against the functionalist thesis. Traditional functionalism primarily emphasises the functional roles that mental states play within an organism. Rupert argues that this kind of functionalism, as traditionally understood, supports the view that cognitive processes are predominantly internal. Rupert challenges

[1] In fact, C&C's 1998 paper does not explicitly mention that functionalism is required to support the EM thesis. However, Rupert identified functional poise in the spread of epistemic credit and argued against it. Thereafter, Clark (2008) developed an explicit requirement for functionalism to support EM's need for a constitutive role for an external artefact in cognition. Following Clark, Wheeler (2010) expanded extended functionalism and addressed the concerns raised by Rupert, A&A, Rowlands and Sprevak.

The historical development of the parity principle and functional poise can be summarised as follows:

1. Spread of epistemic credit (C&C, 1998)
2. Rupert's challenges to functional poise (Rupert, 2004)
3. Parity principle and functional poise (Clark, 2008)
4. Extended functionalism (Wheeler, 2010)

The core of the parity principle and functional poise originated from the requirement for the spread of epistemic credit.

functionalist poise in EM and concludes that traditional functionalism supports internal cognition and cannot be used to support extended cognition.

The internalist criticism of the parity principle is based on the differences between biological processes and external artefacts in cognition. Although the parity principle is not about the similarities between biological cognitive processes and external artefacts, the basis of the internalists' criticism is about such similarities, since biological processes are distinct from external artefacts. The criticisms relate to the requirement and nature of functionalism (coarse- versus fine-grained) for the EM thesis, the portability and availability of biological cognitive processes versus external artefacts, and the mark of cognition (such as non-derived content compared to derived content).

Another issue relates to perception and action. Perception acts as the boundary of mental processes and action is about how the mind acts on the environment. Here, cognitive processes are internal to the agent. Thus, the existence of this boundary is not consistent with P2 and P3, i.e. with causal coupling and the constitutive role of external artefacts. According to Hurley (1998), this tendency to view perception as input from the world to the mind and action as output from the mind to the world persists. Hurley argues that it is wrong to view perception and action as separate types of events. As such, externalism rejects this input–output conception of the mind, which leads to the impression that the mind is separate from the world. Hurley (1998, p. 214) suggests that a change in our conception of this process is required to take us from "the input–output model of perception" to "a two-level interdependence picture" (of "dynamic feedback loops").

Rupert and A&A raise concerns against P3, as they claim that internal cognitive processes are fundamentally different from processes relying on external artefacts, just like internal memory differs from external memory. For example, internal memory suffers from negative transfers and there is the generation effect, neither of which occur in external memory.

Menary (2010) argues that C&C's formulation of the EM thesis, i.e. first-wave EM, relies heavily on parity and functionalism, which results in a misconception that internal cognitive processes resemble external cognitive processes. Menary agrees that the EM thesis is based on the functional roles of internal and external cognition and not on the similarity of external and internal cognition. He supports the second-wave EM proposed by Sutton, in which inner and outer cognitive processes provide a complementary role rather than parity in functional roles. According to Sutton, the functional roles of inner and outer cognitive processes can be different; however, they can complement each other in the overall cognitive process. Menary (2010) states that:

> Cognitive integration takes the first wave of EM arguments to establish that cognition is hybrid. However, it is not motivated by the parity principle, but rather takes embodied engagement with the world as its starting point. The manipulation thesis provides a further motivation and a definition of integrated cognition.

Although Rowlands was a staunch proponent of the EM thesis, he was sceptical about parity and functionalist poise. Sutton disagrees with the parity principle as it "threatens to flatten out the important differences between cognitive artefacts" (Menary, 2010, p. 199). I argue that in C&C's formulation of EM, the constitutive role of an external artefact is established by the parity principle and functionalism. In the EM thesis, an external artefact can be part of an agent's mind if the external artefact plays a causal functional role in accomplishing a cognitive task.

2.3.4 Rupert's Challenge to Parity and Functionalism

Next, I consider Rupert's (2004) challenge to EM, Clark's responses, Martin Wheeler's attempt to extend functionalism and Martin Wheeler's replies to Rowland's deadlock and Sprevak's dilemma. Rupert challenges the functional poise in EM due to the dissimilarities between internal and external cognitive processes. According to Rupert (2004, pp. 421–422): "Clark and Chalmers do not present their position as an explicit development of the functionalist program in philosophy of mind." The functional poise in the spread of epistemic credit was reconstructed by Rupert as follows:

Premise 1: A mental state of kind F is realised by whatever physical state plays the functional role that is characteristic (or metaphysically individuative) of F.

Premise 2: Some realisations of functional mental states have physical components external to the organism.

Premise 3: A mental state extends to or includes all components of its realisation.

Conclusion: Therefore, some mental states extend beyond the boundaries of the organism.

Rupert asserts that if the functional role of an internal cognitive state differs from the role of the relevant external artefact, then premise 2 will fail. Consider the functional roles of internal and external memory, like Otto's notebook. Rupert argues that an internal memory can be accessed instantaneously without conscious thought and may be deeply interconnected with other memories. In contrast, an external notebook requires deliberate action to access and may not be integrated with other memories. Moreover, it depends differently on sensory and motor processes. Rupert argues that internal cognitive processes are more reliable, faster and better integrated. Recalling a memory from the brain is much faster and more reliable than finding information in a notebook. Thus, he shows that functionalism does not support the EM thesis.

When Rupert (2004) challenged the functionalist strain of the EM thesis, no specific functionalist account had yet been provided by C&C for the EM thesis. In evaluating the EM thesis, Rupert considers two predominant views of functionalism: (1) analytical functionalism and (2) psychofunctionalism. Analytical functionalism explains the meaning of mental states, such as being in pain or the belief that it is snowing outside. There are no empirical claims in analytical functionalism. Rupert argues that an analysis of common-sense psychological concepts should yield functional-role descriptions of mental or cognitive states, such as memory. Rupert considers memory as a cognitive state to evaluate whether analytical functionalism supports the EM thesis. Rupert states that an analysis of the common-sense psychological concept memory yields a functional description as "a memory that P is, among other things, caused by interaction with a certain state of affairs (which we might normally describe as the content of P or what the memory that P is a memory of) and, under certain conditions, a cause of the belief that P" (Rupert, 2004, p. 422). It seems that Rupert's functional description of memory indicates that there is a possible conflation with the concept of belief and a lack of emphasis on the unique features of memory as a cognitive process. Note that accessing a memory typically involves recalling information about past experiences, whereas a belief can be based on current information or reasoning.

Rupert (2004) argues that such an analysis of memory does not support HEC. Even if Otto relies on his notebook in the same way that others rely on their biological memory, our common-sense understanding of memory does not encompass external entities. Saying that Otto "remembers" an address by looking it up in his notebook feels like a misuse of the term "remember" according to our everyday language and understanding. When we think of someone remembering something, we typically imagine them accessing some internal database in their brain. In Otto's case, this "database" is his notebook. Rupert argues that this runs counter to our intuitions. We do not typically conceive of external objects, like notebooks, as being part of our cognitive apparatus, no matter how reliant we are on them. It is counter-intuitive to say that I saw a memory of yesterday's trip somewhere. "For common sense rules strongly against external portions of memories and other cognitive states. The common-sense conception of memory precludes its being seen by its possessor" (Rupert, 2004, p. 422). Rupert argues that: "The analysis of common-sense concepts of cognitive states does not support the hypothesis of extended cognition, for common sense rules strongly against external portions of memories and other cognitive states" (Rupert, 2004, p. 422)". Rupert suggests that, according to functionalism, it is counter-intuitive to consider that our own memory resides in a notebook, as it does for Otto.

Rupert (2004) then considers psychofunctionalism, which attempts to explain the nature of mental terms empirically. It asserts that psychological states are characterised by their inputs and outputs and the states that connect them. In psychofunctionalism, the characterisation of the individuating functional roles of mental states is given by our best psychological theories. Rupert argues that the functionalist approach to HEC is inadequate, as it should be able to explain and predict behaviour. External artefacts are not considered in psychofunctionalism. Rupert (2004) argues that cognitive science is explanatory. Internal memory has various properties, including: (1) the generation effect, i.e. information is remembered better if it is generated by the agent rather than being learned, and (2) the recency effect, i.e. the tendency to remember recent information better. Biological memory has these complex characteristics that cognitive scientists study and try to explain, whereas an external artefact, like a notebook, does not inherently have these features. A notebook does not "forget" over time or show a generation effect. It remains static, storing information without the nuances and intricacies of biological memory. By considering the differences in how internal and external memories operate, Rupert argues that they serve different functional roles. Even if both store information, the manner and the complexities with which they do so differ significantly. Therefore, the functional role of internal memory is different from the functional role of external memory.

The main claim of HEC is that cognitive processes can extend beyond the individual's brain and even their body to include parts of the external environment. Here an external artefact has a constitutive role, i.e. it is an integral part of the cognitive system. Rupert's counterproposal is the hypothesis of embedded cognition (HEMC). According to HEMC, external tools and artefacts are certainly influential and play essential roles in our cognitive processes, but they are not considered part of the cognitive process itself. Instead, these artefacts are seen as being causally effective. They influence cognitive processes but are not a constitutive part of those processes. In this perspective, a notebook used to remember something is just a causal aid to memory, not part of the memory process itself. Rupert claims that even if HEC theorists argue that generic memory has an internal part and an external part, his formulation of embedded cognition (HEMC), i.e. that an external artefact has a causal role rather than a constitutive role, can better explain such roles. According to Rupert, external memory, such as a notebook, allows the agent to remember but it is not part of the actual memory process. Therefore, Rupert concludes that functionalism does not give independent support to EM.

Using fine-grained psychofunctionalism, Rupert argues that there need to be similarities between internal and external processes if the external part is to be considered as cognitive. Clark argues that EM requires only coarse-grained common-sense functionalism, for which similarities between internal cognitive processes and external processes involving an artefact are not required.

Clark's (2008) response was that the EM thesis does not require there to be a similarity between the inner and outer parts of the mind; therefore, there is no requirement that all systemic elements must behave according to the same laws. Clark argues that to assume that the outer parts of the system follow the same laws as the inner parts begs the question of whether the target of cognitive science is a hybrid system with inner and outer elements. This would necessitate additional principles governing the overall hybrid mind.

Clark (2008, p. 88) argues that the key issues concern coupling only indirectly; what matters is the functional poise achieved for the stored information. According to this view, it does not necessarily matter where the belief is stored, whether in a notebook or in the brain. What matters is the functional role. Instead of getting into the specifics of where information is stored or how tightly an agent is coupled with an external artefact, the focus should be on how the information functions in guiding thought and action. A necessary and sufficient condition for C&C's version of EM is that in accomplishing a cognitive task, an external artefact has a causal functional role. However, the dynamic interaction of the agent with the external artefact is neither necessary nor sufficient for EM.

There are arguments for the incompatibility of functionalism and EM. As Rupert argues, the functional roles played by external artefacts are fundamentally different from those played by internal mental states. As we have seen earlier, traditional computational functionalism does not necessarily posit a continuous interplay between internal states and an external artefact. The EM thesis emphasises this continuity, as cognition can flow seamlessly between brain, body and environment. Although, there are arguments against the incompatibility of functionalism and EM, it is not necessarily the case that no forms of functionalism can support EM. The compatibility depends on the specifics of the functionalist position in question.

2.3.5 Michael Wheeler's Formulation of Extended Functionalism

Wheeler (2010) follows Clark and argues that the logical development of functionalism is extended functionalism. Wheeler argues that instead of the traditional functionalist approach to sensory input, behavioural output and mental states, as Levin (2008) formulates, what makes a systemic state a mental state is the set of causal relations that it bears to systemic inputs, systemic outputs and other systemic states. Wheeler claims that this formulation of functionalism can expand the border of cognition beyond sensory inputs and allow an external artefact to have a constitutional role in cognition. Wheeler terms this extended functionalism. Wheeler's main thrust was to rely on multiple realisability, which supports coarse-grained functionalism.

As noted above, Rupert's strategy was to highlight the difference between cognitive traits in internal (biological) memory and in external memory to show that the functional roles of internal and external memory are different. Thus, the functional poise in HEC fails to support the constitutional role of external artefacts in cognition.

Wheeler's strategy was to make the features of external memory the same as those of internal memory to demonstrate that Rupert's concern about the difference between biological and external memory is, in fact, due only to their spatial location, i.e. one is inside the skull and skin bag (internal memory) and the other is external to the skin bag. Wheeler's appeal to multiple realisability suggests that just as a particular cognitive function can be realised in different kinds of biological systems, it could also be realised in a non-biological system.

Wheeler argues that Rupert relies on fine-grained chauvinistic functionalism, which requires internal and external memory to be similar. Wheeler argues that the functionalist poise of EM does not require fine-grained functionalism. Instead, it needs only coarse-grained functionalism, such as generic memory, in which external and internal memory can have hybrid functional roles in cognitive processes. Internal memory has various properties, including the generation effect and recency effect. External memory has neither of these properties. However, both internal and external memory can play a hybrid role in achieving a cognitive task. If we accept that the same cognitive function can be realised in many ways, both internally and externally, then we should adopt a functionalist perspective that can accommodate this variability. A coarse-grained functionalist perspective is necessary for a hybrid generic memory. This is because the exact ways in which internal and external components interact can vary widely across different cases, and a fine-grained approach cannot accommodate all these variations. Rupert, in contrast, insists on a fine-grained approach in which external components cannot be genuinely cognitive. Wheeler argues that if a coarse-grained perspective allows for the possibility of extended cognition and supports the principle of multiple realisability, then why not adopt it? In essence, Wheeler's argument is that if we genuinely want to explore the possibilities of extended cognition and do justice to the principle of multiple realisability, then a coarse-grained functionalist perspective is not only preferable but, perhaps, necessary.

Rowlands (2010a) notes that internalists can raise a similar question. Why should we adopt a coarse-grained functionalism that supports extended cognition when a more fine-grained approach that focuses solely on internal cognitive processes would suffice? Wheeler argues that according to multiple realisability, the same cognitive function could be instantiated in various physical systems or configurations. A coarse-grained approach is inherently more open to these multiple realisations, whether they are within the brain, across different biological entities or even external, non-biological systems. Wheeler describes empirical findings of an external artefact integrated into an individual's cognitive processes such that they function as internal processes, which serves as evidence favouring a coarse-grained approach. Cognitive psychologists are interested only in the functional role of the memory, rather than whether it is fine-grained or coarse-grained or whether it has a generation effect. Thus, Rowlands's deadlock fails.

A&A (2010) argue that a high-level cognitive process can be realised by distinctive lower-level processes. A&A agree in principle that some cognition-realising substrates may involve (wholly or partly) non-neuronal elements. However, they reject this for humans and argue that for humans, all lower-level processes should be neuronal. This contradicts the multiple realisability of functionalism for the mind.

Sprevak (2009) raises concerns about the coarse-grained functionalist requirement for EM based on a Martian's intuition. Sprevak argues that Martian intuition is possible if functionalism is coarse-grained rather than fine-grained. However, coarse-grained functionalism in HEC with the parity principle can result in an unwanted extension of cognition. This leads to a dilemma: either functionalism is false or an implausible version of HEC is true.

Wheeler identifies three conceptual factors in Sprevak's dilemma: (1) a functionalist understanding of extended cognition, (2) the independent plausibility of the Martian's intuition and (3) the centrality of the parity principle for extended cognition.

Sprevak's strategy has the following steps:

Step 1: Postulate a HEC example with an external artefact, like Otto and his notebook, or a cognitive aid, like a computer and the Mayan calendar program.

Step 2: Imagine a functionally identical system located entirely inside the head of a Martian. As in HEC, the Martian's intuition works like the external artefact.

Step 3: Apply the parity principle to the HEC example of step 1 and check whether the external artefact is part of a cognitive process. For Otto, his notebook is cognitive.

Step 4: Either we must deny that the Martian has beliefs matching the contents of his internal flesh-book or admit that Otto has extended beliefs matching the contents of his notebook.

Wheeler argues against step 2, i.e. something inside the Martian's head that resembles a non-cognitive external artefact, and he suggests that this thing is not cognitive. Wheeler argues that whether the thing is cognitive is not based on its spatial location. Therefore, Wheeler rejects the Martian's intuition in step 2.

Wheeler also argues against step 3, which uses the parity principle, based on the location of the artefact. Wheeler argues that the "parity principle should have *equal treatment regardless of location*. Thus, the parity principle also implies that an external element that we take to be non-cognitive doesn't become cognitive *purely in virtue of being shifted inside the head*."

2.3.6 Requirement of Functionalism for EM Thesis

The debate over EM and functionalism revolves around whether cognitive processes and states can extend beyond the brain and into the external world. Functionalism, which posits that mental states are defined by their functional roles and not by their underlying physical substrates, seems a natural fit for arguments in favour of EM. The arguments for applying functionalism to EM are as follows:

1. Multiple realisability: One of the cornerstone ideas of functionalism is that a particular mental state can be realised in multiple ways, not just in biological neural networks but potentially in other substrates too. This idea naturally lends itself to the claim of EM, which suggests that cognitive processes can be realised in external artefacts, if they perform the right function.

2. Functional equivalence: Proponents of EM argue that if an external artefact, like a notebook, plays a functionally equivalent role to some internal cognitive process, like memory, then it should be considered a part of the cognitive system.

3. Empirical observations: There are many real-world examples in which an external artefact seems deeply integrated into human cognitive processes. Functionalism provides a framework to describe these cases without committing to a specific physical realisation of the cognitive process.

However, as seen earlier, there are arguments against using functionalism as a foundation for extended cognition. Arguments against applying functionalism for EM are as follows:

1. Overly liberal criteria: Critics like Sprevak argue that a coarse-grained functionalist approach leads to an overly broad form of EM. If any external artefact or process that plays a role that is functionally equivalent to an internal process can be seen as cognitive, then almost anything could be considered an extension of the mind, leading to seemingly absurd conclusions.

2. Undermining intuitive cases: A fine-grained functionalist approach, on the other hand, could set criteria for EC that are too restrictive. These criteria could exclude many cases that proponents of EM would intuitively want to count as instances of EC.

3. Inherent nature of cognitive states: Some argue that even if an external tool plays a functionally equivalent role, it does not have the inherent nature of certain

cognitive states. For instance, a notebook may help in memory retrieval, but it lacks the experience associated with remembering.

4. Causal versus constitutive role: Critics like Rupert argue for a distinction between an external artefact playing a causal role in cognition versus one having a constitutive part of a cognitive process. Although functionalism could capture the causal role, it may not adequately address the constitutive aspect.

In summary, although functionalism offers a seemingly compatible framework for EM, the specifics of how functional roles are defined and the consequences of those definitions have not yet been settled.

I argue that even extended coarse-grained common-sense functionalism is not consistent with the EM thesis. Wheeler's attempt to replace sensory input, mental states and output beliefs by systemic input, systemic belief states and systemic output does not change the core internalist aspect of functionalism because it is still based on relations between inputs and outputs without considering dynamic interactions. I argue that neither functionalism nor extended functionalism based on parity is required for the EM thesis. The reasoning for my argument can be summarised as follows:

1. Traditional functionalism rests on an internalist view of the mind that is consistent with CTC.
2. In C&C's EM thesis, an external artefact can be part of an agent's mind if the external artefact plays a causal functional role in achieving a cognitive task. The dynamic interaction of the agent with an external artefact is neither necessary nor sufficient for EM.
3. Extended functionalism with systemic inputs and outputs can incorporate the environment in the cognitive processing of an agent. However, the sequence (the order of the inputs and outputs in serial computing like CTC) is important for functionalism. The dynamic engagement of the agent with the environment is neither necessary nor sufficient for extended functionalism.
4. Wheeler (2010) states that traditional functionalism also faces philosophical challenges connected with phenomenal consciousness, the "what-it's-like-ness" of experience: "Extended functionalism inherits the disadvantages, as well as the advantages, of its parent theory."
5. Extended functionalism is not required in proving the externalist aspect of the EM thesis. This is demonstrated by the vehicle externalism proposed by Hurley

(Hurley's how-enabling vehicle externalism about phenomenal qualities and labelling EM as how-enabling content externalism).

6. Various EM examples provided by Clark, such as Otto, Ada, Addler, the dancer etc., have a varying degree of constitution with the external surroundings, and they do not necessarily require extended functionalism. Therefore, functionalism is not required by the EM thesis. Parity is not sufficient for EM. Both dynamical coupling and information processing have necessary roles in constituting whether an external artefact is involved in cognition.
7. My conclusive argument is that functionalism based on parity is neither necessary nor sufficient for EM.

2.4 Clark's Position on Various Versions of EM

The first part of this section evaluates Clark's position on anti-Cartesian cognitive science and the second part evaluates various versions of EM to explore the potential for developing the EM thesis without functionalism to address the coupling-constitution fallacy and cognitive bloat.

Clark's position is explained very well by Chalmers:

> The extended mind thesis is compatible with both physicalism and dualism about the mental. It is compatible with connectionist and classical views, with computational and noncomputational approaches, and even with internalism and externalism in the traditional debates over mental content (as we suggest in *The Extended Mind*). So I do not think that the extended mind thesis requires much in the way of theoretical presupposition at all. Instead, it is an independently attractive view of the mental.
>
> (Clark, 2008, Foreword, pp. XV and XVI)

Chalmers also argues that functionalism is not a strict requirement for the EM hypothesis: "The deepest support for the view comes from the explanatory insights that the extended mind perspective yields." Chalmers emphasises the explanatory power of EM to explain mental processes.

There are differences in the approaches to EM. Not everyone agrees that the EM thesis requires external artefacts to have a causal functional role based on parity. According to Clark (2008) and Wheeler (2010), the extension of the mind via an external artefact is possible if it has the right kind of causal or functional role. Menary (2010) proposes an alternative version of EM based on cognitive integration, in which an extension of the mind is possible when the external artefact and mental processes involve bodily manipulation, i.e. manipulation of information-bearing structures to gather information for the accomplishment of a cognitive task. In this approach, cognitive processes are hybrids of internal and external cognitive activities. Sutton (2010) argues that functionalism is not required for EM and an extension of the mind is possible when the external artefact has a role that is complementary to the inner biology.

Clark considers that only some mental processes, such as beliefs, can be extended through an interaction with cultural artefacts. Consciousness cannot be extended, as it is wholly internal to the cogniser. Likewise, qualia cannot be extended, as they too are internal to the agent. Only beliefs derived from cognitive processes can extend into the environment. However, Rowlands (2010b) argues that consciousness can be extended through interactions between an external information-bearing structure and the cogniser.

Rowlands (2010a) proposes an alternative version of EM with a mark of cognition. Hurley suggests quality-enabling externalism in which phenomenal qualities can be extended by the interactions of the cogniser with cultural and non-cultural artefacts.

The approaches suggested by Rowlands and Hurley for how the mind can be extended are quite different from Clark's. However, there is not much difference between Sutton's and Menary's versions of EM and Clark's version, except for the lack of dependence on both functionalism and the parity principle. Note that Rowlands, Sutton and Menary consider the EM thesis to be an ontic thesis, such that the constitutive role of an external artefact is essential for establishing the metaphysical claim of the EM thesis. Hurley, however, does not consider the metaphysical requirement for the EM thesis, but considers only the merits of its explanatory potential.

Although C&C's version is anchored on common-sense functionalism, i.e. coarse-grained functionalism, Menary's approach emphasises the transformation of cognitive processes through cultural practices and tools. By focusing on how tools and practices deeply transform and are integrated into cognitive tasks, it can be seen that not everything in the environment can be cognitive. For a cognitive process or artefact to be constitutive, it needs to be integrated in a way that transforms cognitive tasks. This argument is a defence against cognitive bloat and the coupling-constitution fallacy. Although Menary's version emphasises cognitive integration, it does not clearly define what it means for an artefact or a cognitive process to be integrated into cognitive routines. For example, a pen can be deeply integrated into the processes of academic thought but that does not necessarily make it a part of cognition. It is debatable in such cases whether the artefacts themselves become parts of cognition or remain as aids to cognitive processes. There are challenges in clearly demarcating what counts as genuinely cognitive and what does not. Therefore, there are the potential pitfalls of cognitive bloat and the coupling-constitution fallacy. Sutton's second-wave EM approach is based on the complementarity principle. Although Sutton rejects parity and embraces complementarity instead, second-wave EM has no specific advantage compared with first-wave EM in terms of countering criticisms of it, such as the coupling-constitution fallacy, the enabling versus constitutive role of external artefacts and cognitive bloat. Therefore, second-wave EM based on complementarity is vulnerable to objections against EM.

After evaluating all these versions of EM, I conclude that C&C's original proposal for EM without functionalist poise has the potential to develop in new directions compared with the other versions. Therefore, my aim is to modify C&C's original EM thesis to make it immune to the coupling-constitution fallacy and cognitive bloat.

2.5 Conclusions

I conclude that EM gives a better perspective of cognition than the traditional internalist account. The argument against EM based on the internalist account of cognition, i.e. that the bounds of cognition are inside the skin and within the skull, ultimately fails to undermine EM. I also argue that extended, embodied and integrated cognition can better explain cognition than the traditional internalist account in which cognition is internal to the agent.

I conclude that with Clark's formulation of EM, it is difficult to establish the externalist claim that the mind extends into the world. It explains the role of external artefacts in cognition via a causal or dependent role, as in HEMC. Clark's reliance on functionalism is difficult to reconcile with the co-stream of cognitive thinking, such as embodied and enactive cognition. Traditional functionalism is neutral with respect to cognitive processes and accepts CTC, such that the brain is like a syntactic and semantic machine with internal representation. However, embodied, dynamic and enactive cognition rejects this serial computing aspect of cognition.

Establishing criteria for the mark of cognition and interpreting whether mental states meet those criteria does not have open and unprejudiced research potential. The criteria may not have a role in establishing the constitutive or causal role of external features in cognition and mental states.

The explanatory power to account for the historical progress of patterns and the abstraction of thought processes and its impact on the development of tools, artefacts, social institutions and cognition should be considered as important in establishing the role of external features in cognition and mental states.

Proponents of EC, such as Rowlands, Menary and Sutton, oppose the common-sense functionalism employed by C&C. However, as explained above, these alternative versions of EC still struggle to address the coupling-constitution fallacy and cognitive bloat. Considering the potential of DST, as in the analogy of a Watt governor to cognition, a framework for EC based on DST could potentially overcome issues like the coupling-constitution fallacy and cognitive bloat without necessitating the controversial stance of functionalism. This is the focus of the next chapter.

3 Modified and Integrated Version of the EM Thesis

3.1 Introduction

This chapter focuses on modifying the Clarkian version of extended mind (EM) to make it immune to the coupling-constitution fallacy and cognitive bloat. Once modified, I attempt to integrate the modified EM with niche construction theory (NCT), cognitive niche theories and developmental systems theory to establish the extent and boundaries of EM and cognitive development and the nature of a cognitive agent.

After evaluating the criticisms and various versions of the EM hypothesis, I conclude that, in Clark's formulation of EM, it is difficult to establish the constitutive role of an external artefact in cognition. Rather, an internalist can reconcile the importance of an external artefact in cognition, such as embedded cognition, in which external artefacts have only an enabling role rather than a constitutive role.

As can be seen in Chapter 2, the criticisms of the enabling versus the constitutive role of an external artefact in cognition and cognitive bloat put some pressure on Clark and Chalmers's (C&C's) version of EM. Clark's reliance on functionalist theory (i.e. cognition understood based on functional roles) and explained with the parity principle (i.e. the equivalence of the role of internal cognitive functions compared with the functional roles of an external artefact in cognition) was not sufficient to counter these criticisms. Moreover, functionalist poise clearly deviates from the core of four E's cognition (embodied, enactive, embedded and extended cognition), i.e. against the Cartesian notion that the boundary of the mind is inside the skin and skull.

I have formulated a version of Clarkian EM in which the externalist criteria for EM can be explained by feedback loops involving external artefacts. I am going to argue that the existence of a feedback loop with an external information-bearing structure and the manipulation of the external information-bearing structure are a necessary and sufficient condition for EM. This modified version of EM is consistent with embedded, embodied, dynamic and enactive cognition and immune to criticisms like the coupling-constitution fallacy and cognitive bloat and issues with the mark of cognition.

This chapter mainly focuses on the necessary and sufficient conditions for extended cognition (EC) based on dynamical systems theory (DST). It applies various arguments from developmental systems theory and NCT and an argument from the patterns and historicism embedded in external information-bearing structures and the role of a pattern recogniser. This chapter also explores the integration of EC and NCT to explain how intellectual abilities arise from the innate cognitive abilities of humans endowed by evolution.

Section 3.2 describes the universality of feedback loops in the environment. Section 3.3 focuses on arguments from DST relating to cognition. In Section 3.4, I detail developmental systems theory. Section 3.5 gives arguments from NCT regarding the feedback loop between an agent and the external environment. Section 3.6 argues from the history of patterns and patterns embedded in external information-bearing structures. In Section 3.7, I detail the modified version of EM and critically analyse the various examples of EM and the criticisms of EM. I conclude that the modified EM is immune to the coupling-constitution fallacy and cognitive bloat. Section 3.8 focuses on the scope of integrated EM. It notes the versatility and potential of EM based on NCT, DST and developmental systems theory. In the same section, I describe the role, extent and boundaries of the hypothesis of extended cognition (HEC) in the big picture of cognition, which is based on the broad patterns from integrated EM. EM is a philosophical theory of the mental whereas NCT is a scientific approach in evolutionary biology. I conclude that EM, DST, NCT and developmental systems theory reveal the inherent potential of EM to explain the cognitive development and nature of a cognitive agent. Integrated EM is useful in a wide range of domains and is not limited to being a philosophy of mind. It has the potential to support an integrated approach to multidisciplinary research.

3.2 Feedback Loops with External Information-bearing Structures

In Chapter 2, I summarised the following.

1. Clark’s version of EM clearly has inherent contradictions due to its dependence on functionalist theory to prove the constitutive role of an external artefact in cognition.

2. The hypothesis of embedded cognition (HEMC) raises serious concerns about the requirement for external artefacts to have a constitutive role in cognition. Clark’s responses are not wholly successful in establishing such a constitutive role.

3. The trust and glue conditions are not sufficient to avoid cognitive bloat. Even for the counter-intuitive cases, it is easy to satisfy the trust and glue conditions such that any external artefact can be considered as part of the cognitive system.

4. The various versions of the EM thesis clearly have similar flaws. None of these alternative versions of EM can successfully establish the constitutive role of external artefacts and avoid cognitive bloat.

My aim here is to formulate a modified version of EM that is immune to the criticism of the enabling versus constitutive role of external artefacts. To retain the original spirit of the EM thesis, i.e. as an anti-Cartesian cognitive science, I argued in Chapter 2 that it is necessary to avoid relying on functionalism to support the constitutive role of an external artefact in cognition. To modify EM to make it immune to the coupling-constitution fallacy and cognitive bloat, I rely on feedback loops between the agent and external information-bearing structures and the manipulation of the external information-bearing structures by the agent as the basis of a modified EM supported by DST, developmental systems theory, NCT, and the patterns and historicism embedded in external information-bearing structures, which I am going to detail in the following sections. Establishing a dynamic feedback loop between the agent and an external artefact and the manipulation of that external artefact by the agent to achieve cognitive success are key for the modified EM. DST establishes feedback loops, which diffuse the dichotomy of mind and action. Based on DST, I can formulate a modified version of EM such that it is immune to the criticism of the enabling versus constitutive role of external artefacts. Hurley (1998) also highlights the importance of a dynamic loop that can diffuse the dichotomy of perception and action.

I explore here various developments, such as the application of DST to cognition, developmental systems theory, NCT from evolutionary biology and cognitive niche theories. Rather than trying to justify the constitutive role of external artefacts in cognition, my strategy is to explore the best explanation for cognition based on various developments in evolutionary biology (NCT), DST, developmental theories, cognitive niche theories and neuroscience and from the patterns and the history associated with external information-bearing structures and pattern recognisers.

This chapter primarily focuses on the development of a modified version of C&C’s EM that satisfies the externalist claim that external artefacts have a constitutive role in cognition rather than the internalist claim that artefacts have an enabling role in cognition. As a result, there is neither cognitive bloat nor a reliance on functionalism. I am going to develop four independent arguments to support the role of feedback loops with external information-bearing structures in cognition:

1. Argument from DST: DST in cognition
2. Argument from developmental systems theory
3. Argument from NCT
4. Argument from the patterns and historicism embedded in external information-bearing structures and the role of a pattern recogniser

The arguments demonstrate the fundamental notion that some cognitive processes have a feedback loop with external information-bearing structures. The manipulation of an external information structure via a feedback loop is a necessary and sufficient condition for EM. Feedback loops are part of the constitutive role of external artefacts in cognition. Note that NCT in evolutionary biology describes the modifications made to an agent's niche through a feedback loop with the environment. DST accounts for the dynamic nature of agents over time.

As Hurley (2010) argues, the EM thesis is about the vehicles of cognitive processes rather than the content. It is about the subpersonal level of explanation for the vehicles of cognitive processes. The arguments in the list above support this explanation of EM based on the vehicles of cognitive processes. The modified version of EM accommodates the core principle of C&C's EM, i.e. the constitutive role of an external artefact in cognition via active externalism and causal coupling.

The argument from the patterns embedded in external information-bearing structures explains the role of these structures in cognition. It demonstrates the versatility of and historicism associated with the EM thesis.

C&C (1998) argue that: "The human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right" (C&C, 1998, p. 8). The close coupling of a cognitive agent with an external artefact is attributed to the feedback loop between the agent and the artefact. However, many loops can possibly form between a cognitive agent and an external artefact. The interactions between cognitive processes and the environment are not limited to feedback loops. Section 3.2.1 explores the various types of loops between a cognitive agent and an external artefact.

3.2.1 Various Loops between a Cogniser and External Artefacts

The nature of the engagement of a cogniser with an external artefact can vary. Some information processing, e.g. the identification of a pattern by the cogniser, may occur in an open loop, i.e. there is no requirement for the agent to manipulate external information-bearing structures when exploring patterns. The offloading of cognitive activities to an external artefact is one form of engagement. Pencil and paper are an example of this. However, to accomplish some cognitive tasks, the manipulation of an external artefact may require continuous reciprocal causation (CRC) between the agent and the external artefact. CRC is based on an ongoing, bidirectional feedback loop in which each entity influences and is influenced by the other. As Clark states, real feedback loops are complex. CRC (as characterised by Clark, 1997a) involves multiple simultaneous interactions and complex dynamic feedback loops, such that: (a) the causal contribution of each systemic component partially determines, and is partially determined by, the causal contributions of large numbers of other systemic components, and, moreover, (b) those contributions can change radically over time. The use of night vision goggles (NVGs) by a soldier is one such example.

To show the differences between feedback loops and open loops, the following sections discuss open and feedback loops in cognitive processes. The sketches in Figure 3.2 and Figure 3.4 show a simple epistemic open-loop system and a simple epistemic feedback-loop system, respectively.

3.2.1.1 Open-Loop Systems

A simple example of an open loop is a toaster (Figure 3.1). Closing the switch toasts the bread at 30°C for 30 seconds. The input to the controller, i.e. 30°C for 30 seconds, is a goal state, and the output from the controller is a signal to the coil to achieve the goal state.

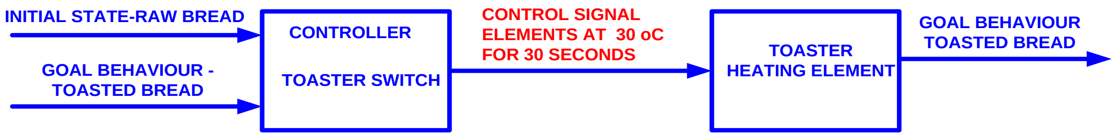
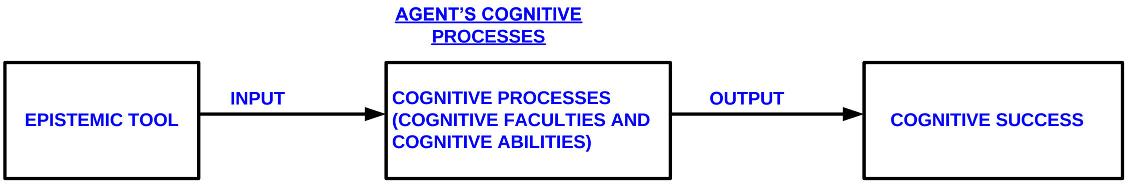


Figure 3.1. Open loop for a toaster.

Figure 3.2 shows a simple epistemic open-loop system and an agent’s cognitive processes without feedback. The epistemic tool is a mere enabler or aid in the agent’s cognitive processes. An example is a thermometer used to measure temperature.


AGENT'S COGNITIVE PROCESSES
EPISTEMIC TOOL
INPUT
COGNITIVE PROCESSES (COGNITIVE FACULTIES AND COGNITIVE ABILITIES)
OUTPUT
COGNITIVE SUCCESS

NOTES

1. NO FEEDBACK LOOPS

2. SINCE THERE IS NO FEEDBACK LOOP FROM THE AGENT TO THE EPISTEMIC TOOL, THE EPISTEMIC TOOL HAS ONLY AN ENABLING ROLE IN THE COGNITIVE SUCCESS AND NOT A CONSTITUTIVE ROLE

Figure 3.2. Simple epistemic open-loop system and an agent's cognitive processes without feedback.

3.2.1.2 Feedback Loops

A feedback loop is a closed-loop system (Figure 3.3). A simple example of a system with a feedback loop is a thermostat switch that maintains the temperature of a room at a set point, say 18°C. The room temperature is fed back to the controller. Whenever the room temperature falls below 18°C, a signal is sent to a heater to achieve the required room temperature of 18°C. The feedback loop works with an error signal. That is, the room temperature, which is the goal state, must have deviated from the set point before the controller can respond (Morris & Langari, 2012, p. 9).

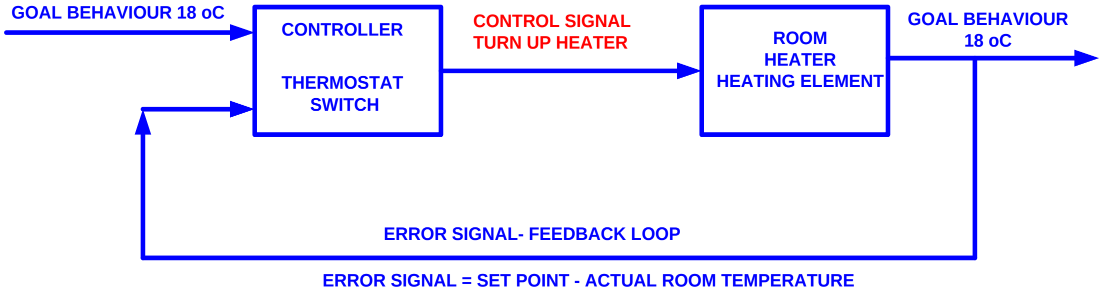
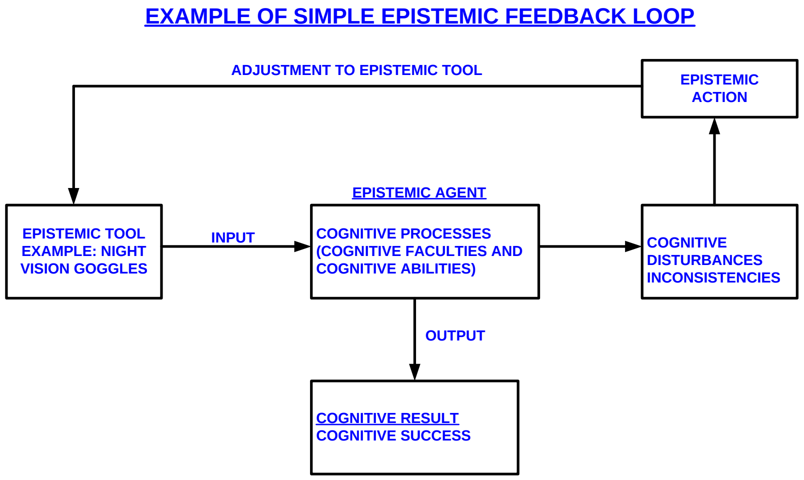

GOAL BEHAVIOUR 18 oC
CONTROLLER
THERMOSTAT SWITCH
CONTROL SIGNAL
TURN UP HEATER
ROOM HEATER HEATING ELEMENT
GOAL BEHAVIOUR 18 oC
ERROR SIGNAL- FEEDBACK LOOP
ERROR SIGNAL = SET POINT - ACTUAL ROOM TEMPERATURE

Figure 3.3. Example of a feedback loop.

A simple epistemic feedback loop is shown in Figure 3.4. As can be seen, an external artefact can have a constitutive role in a feedback loop, and thus, in the overall cognitive processes of the agent. An example of a simple epistemic feedback loop is the use of NVGs.


EXAMPLE OF SIMPLE EPISTEMIC FEEDBACK LOOP
ADJUSTMENT TO EPISTEMIC TOOL
EPISTEMIC ACTION
EPISTEMIC AGENT
EPISTEMIC TOOL EXAMPLE: NIGHT VISION GOGGLES
INPUT
COGNITIVE PROCESSES (COGNITIVE FACULTIES AND COGNITIVE ABILITIES)
COGNITIVE DISTURBANCES INCONSISTENCIES
OUTPUT
COGNITIVE RESULT
COGNITIVE SUCCESS

NOTES

1. INTELLECTUAL VIRTUES ARE REQUIRED FOR THE INITIAL SET - UP OF FEEDBACK LOOPS

2. EPISTEMIC TOOL HAS A CONSTITUTIVE ROLE.

3. EPISTEMIC TOOL WILL BE A CONSTITUTIVE PART OF THE OVERALL COGNITIVE PROCESSES, ONCE IT IS PROPERLY INTEGRATED WITH THE AGENT'S COGNITIVE CHARACTER IN A SEAMLESS FASHION AND IT BECOMES AN EXTENDED COGNITIVE ABILITY. IN SUCH SITUATIONS, INTELLECTUAL VIRTUES WILL NOT BE IN OPERATION AND THE AGENT CAN COGNISE UNREFLECTIVELY.

Figure 3.4. Simple epistemic feedback-loop system and an agent's cognitive processes with feedback.

Clark and Grush (1999) argue that this – feedback-oriented – approach is inadequate for many types of real-world interaction, as such interactions are numerous and complex. Clark (1997a) describes the complexity in the neural circuitry of robots. Clark doubts that we could understand such complex systems, especially the human brain:

> Fans of real-world robotics note that researchers routinely underestimate the difficulty of problems (by ignoring such real-world features as noise and the unreliability of mechanical parts) and also fail to spot quick and dirty solutions that depend on such gross physical properties as the elasticity and "give" of certain parts. (Clark, 1997a, p. 95)

Clark states that "even the 30-neuron leg controller constitutes a dynamical system of such complexity that our intuitive geometric understanding breaks down" (Clark, 1997a, p. 101). Therefore, using DST to gain an understanding of the neural circuitry underlying human behaviour will be extremely complex.

3.2.1.3 Various Loops and the Brain

Those philosophers who have not accounted for developments in neuroscience in their philosophy of mind were nicknamed "non-brainers" by Churchland (2002). I agree with

Churchland that any mature philosophy of mind should incorporate the findings of neuroscience. Current developments, for example, suggest that brain cells can function simultaneously in multiple loops when accomplishing brain activities. The brain has multiple simultaneous loops, with various types of loop and interactions, such as open loops, feedforward loops, feedback loops, etc.

Rather than Grush's emulation model, which is based on a feedforward loop from control theory (Grush, 2004), I utilise another model from control theory (Figure 3.5). For cognitive processes, I suggest that master and slave control is used to achieve the goal state. It is also called cascade control. In this system, which includes two controllers and two measuring elements, the output of the master controller is used to adjust the set point of a slave controller (Coughanowr, 1991, p. 250). In a cascade loop, the goal behaviour is primarily controlled by a slave controller. When the load changes on the manipulated variable, the master controller will take control. The neural basis of a reward circuit can be explained by a cascade loop, which is the neuroscientific basis of addictive behaviour. I use this model to show that brain functioning is not based on single loops but simultaneous multiple loops.

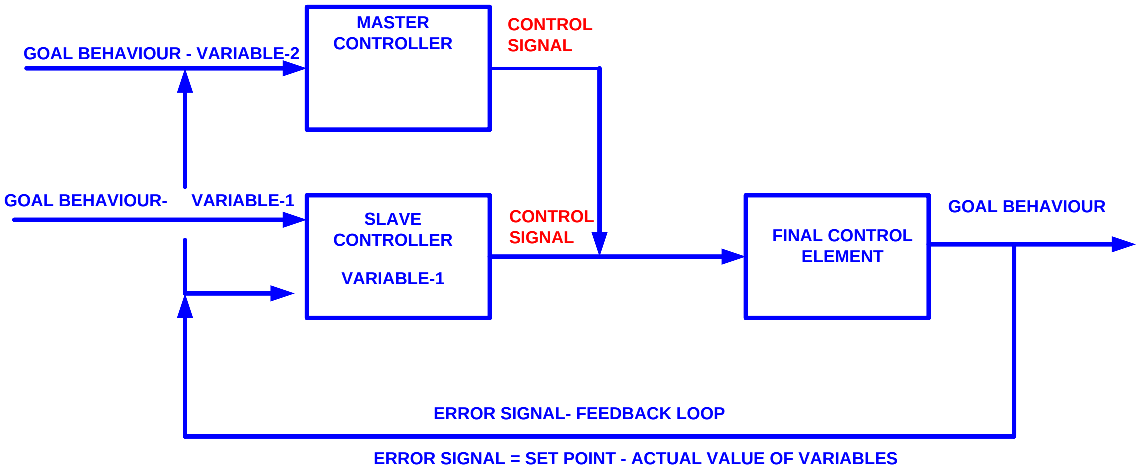


Figure 3.5. Cascade control (e.g. forebrain, nucleus accumbens and VTA).

Under certain conditions, when variable 2 moves beyond the set points, the master controller will exert cascade control over the slave controller, just like the forebrain controls the nucleus accumbens.

As Kalivas and Volkow (2005) assert:

> Neurobiology has focused on three brain regions in the activation of behavior: the amygdala, prefrontal cortex, and nucleus accumbens. The amygdala

> emerged from studies showing involvement in fear-motivated behaviors, while the nucleus accumbens was identified from a connection with reward-motivated behaviors. The prefrontal cortex is less involved in establishing whether a stimulus is positive or negative (valence); rather, it regulates the overall motivational salience and determines the intensity of behavioral responding. (Kalivas & Volkow, 2005, p. 1404)

Figure 3.6 illustrates this circuit. The reward and pleasure system comprises the prefrontal cortex, the nucleus accumbens and the ventral tegmental area (VTA). This cascade loop is also responsible for learning as well as for dangerous addictive behaviours. The nucleus accumbens is the pleasure centre in the human brain. From an evolutionary perspective, it is primitive, as many animals have a nucleus accumbens. The neurons in the nucleus accumbens fire when the VTA releases the neurotransmitter dopamine. The VTA releases dopamine when a surprise or reward occurs. As Kalivas and Volkow (2005) explain:

> Projections from the ventral tegmental area release dopamine throughout the circuit in response to a motivationally relevant event. The release of dopamine signals the circuit to initiate adaptive behavioral responses to the motivational event, and in doing so it facilitates cellular changes that establish learned associations with the event. (Kalivas & Volkow, 2005, p. 1404)

Since the nucleus accumbens is the location of pleasure, it needs to be controlled to avoid harmful behaviour. This is done by the prefrontal cortex, which is the master controller. The prefrontal cortex controls urges and pleasure seeking. The prefrontal cortex is a critical neuroanatomical hub for controlling motivated behaviours across mammalian species. In addition to intra-cortical connectivity, prefrontal projection neurons innervate subcortical structures that contribute to reward-seeking behaviours (Otis et al., 2017, p. 1).

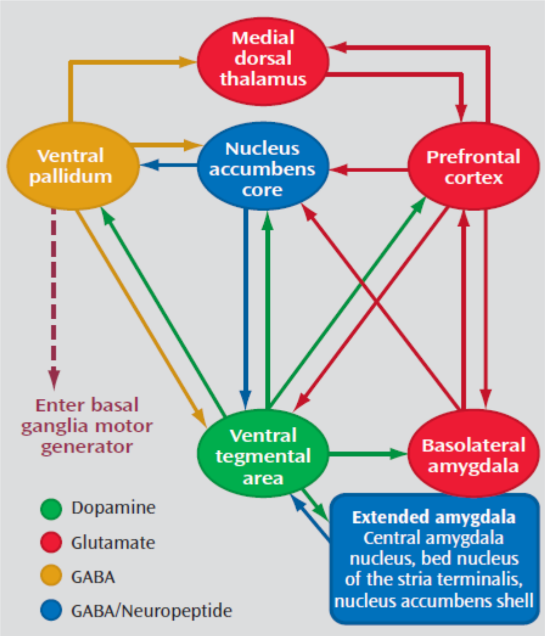


Figure 3.6. Neural circuitry mediating the activation of goal-directed behaviour. Extracted from Kalivas and Volkow (2005, p. 1404).

The important aspect of all sorts of addictive behaviours is associated with the nucleus accumbens, the VTA and the forebrain. Insufficient control by the forebrain can result in addictive behaviour. For example, alcohol blocks the inhibitive neurotransmitter GABA and accelerates the release of dopamine, which stimulates the pleasure centre, the nucleus accumbens.

For example, the consumption of alcohol is popular because of its pleasurable effects during social activities. Alcohol opens the neurotransmitter floodgates. It causes the release of dopamine, serotonin (which governs our sense of well-being) and the brain’s own opioids. It also disturbs the levels of glutamate, which incites neurons to fire and helps account for the initial alcoholic high, as well as GABA, which dampens neuronal firing and eventually makes (most) drinkers sleep (Begley, 2001, p. 40). Once someone is addicted, excessive dopamine released from the VTA results in a craving for alcohol. Over time, their alcohol tolerance will increase. If someone becomes dependent on alcohol and they try to reduce their alcohol intake, they can experience withdrawal symptoms. For all addictive substances, the mechanism is the same: excessive dopamine is released from the VTA, which activates the neurons in the pleasure centre (the nucleus accumbens) and is accompanied by a failure of the master controller (the prefrontal cortex).

The various loops are based on control theory. Many other loops are possible. I argue that cognitive processes must require more than a single loop. They may utilise open loops, feedback loops, feedforward loops or other types of loop like cascade loops. However, my focus here is on a feedback loop involving an agent and external information-bearing structures in cognition.

3.3 Argument from DST: Extended Cognition and DST

The dynamic nature of the cogniser (Beer, 2003; Thelen & Smith 1994) results in cognitive behaviour as the product of interactions among the brain, the body and the world. As described in Chapter 2, the computational theory of cognition does not account for dynamism between the brain/body and the environment. Instead, it considers that all the cognitive activity inside the brain depends solely on computational processes such as symbolic manipulation.

DST describes the dynamic interactions of the agent with the environment and the integrated nature of the body, the brain and the world. Clark states that: “The dynamicist chooses to focus on changes in total state over time” (Clark, 1998, p. 364). Dynamicists have a “strong belief in the relevance of continuity to providing accurate descriptions of cognitive systems” (Eliasmith, 2001, p. 422). As Wilson and Clark (2009) state, various paradigms in the biological sciences, such as NCT, DST and extended physiology, support extended cognition (EC). All these paradigms involve feedback loops.

Palermos proposes a version of Clarkian EM that is immune to the coupling-constitution fallacy and cognitive bloat. Palermos applies DST to cognition to establish the constitutive role of external artefacts in EC. Chemero also responded to the coupling-constitution fallacy using DST.

In this section, I will outline the background and history of DST, its application to cognition and Palermos’s attempt to respond to criticisms of EC by applying DST to cognition as a tool.

3.3.1 Dynamical Systems Theory

Mathematical modelling of an observed system is at the core of the application of dynamical theories. Here is an example of the application of DST in biology. A fictitious ecosystem modelled by Lotka and Volterra (Abraham & Shaw, 1992) contains only two species: big fish and small fish. There is a large supply of food for the small fish, and the big fish eat the small fish. This is an example of a dynamical system with periodic behaviour. In the model, a change in one population will result in a change to the other. If both populations are low, the number of big fish further decreases due to a lack of food (small fish), whereas the number of small fish increases because there is less predation. If there are many small fish and few big fish, both populations will initially increase. The number of big fish increases as they have many small fish to eat. As a result, the population of small fish declines. Eventually, there will be few small fish but many big fish, and both populations will decline. However, as the number of big fish declines, the small fish population recovers. The system is periodic.

The differential equations that describe the path of this trajectory, developed independently by Vito Volterra and Alfred Lotka in the mid-1920s, are fairly simple (reproduced from Shapiro, 2013):

$$\begin{aligned} dx/dt &= \alpha x - \beta xy \quad (\text{prey}) \\ dy/dt &= \gamma xy - \delta y \quad (\text{predators}) \end{aligned}$$

Here the variables *x* and *y* are the number of prey and predators. The Greek letters are parameters that determine the properties of the system. α is the rate at which the prey population would grow in the absence of predators; β is a measure of the rate of predation on prey. Therefore, the first equation describes how the prey population changes over time as a function of its growth rate minus the effect of predation. In the second equation, γ represents the rate of growth of the predator population as a function of the size of the prey population and δ is the rate at which the predator population would die in the absence of prey. Thus, the second equation describes how the predator population grows or shrinks over time as a function of the size of the prey population minus the natural loss of predators.

The advantage of using this mathematical model of this system of big fish and small fish is its predictability. From the model, it is easy to predict the behaviour of the system, i.e. how the numbers of big and small fish evolve over time.

Shapiro (2013) shows that the equations that describe the predator–prey relations also illustrate the idea of coupling. The equations in a dynamical system are coupled. Thus, the rate of change of *x* depends on *y* as well as *x*, and the rate of change of *y* depends on *x* as well as *y*. In this example, how the population of big fish will change depends on the number of small fish, and vice versa. This idea of coupling is fundamental for the claim that cognition extends beyond the brain.

3.3.2 Application of DST to Cognition

DST modelling is a very powerful tool for explaining and predicting natural phenomena, from planetary motion to human biology. The above example of a prey and predator (a small fish and a big fish) is the classical application of DST in biology. It illustrates the periodic increase and decrease of big and small fish over time and the equilibrium. Planetary motion around the Sun is periodic and the position and momentum of each planet can be predicted by DST modelling. “The paths of the orbits of the planets change continuously as a function of the changing relationships the planets bear to each other” (Shapiro, 2013, p. 355).

This success of DST eventually led to the application of DST to psychological activities. The proponents of DST argue that the brain and the mind are parts of the natural world, so logically, they can be modelled by DST. The application of DST to cognition began in 1980 (Shapiro, 2013). In the application of DST, psychological processes and capacities are considered as part of a dynamic system. These psychological processes are complex, non-linear, self-organising and emergent. They develop over the life course of the individual and occur in real time (Spivey, 2007; Van Gelder & Port, 1995). We saw in Chapter 2 van Gelder’s analogy of a Watt governor for cognitive processes.

Systems that continuously interact with the environment, like the brain and body, can be modelled by DST. The psychological and cognitive applications of DST have gained impressive momentum in providing a cognitivist explanation of processes that may be restricted within the brain, processes that extend to the agent’s body and even processes that span the brain, body and the environment (Palermos, 2014a). The cognitivist explanation based on DST modelling can be like finding patterns, as in the works of Kelso, in which the wagging behaviour of fingers could be identified as a pattern in such a way that the movement of fingers could be predicted (Kelso, 1995). However, Beer uses a neural network to explain cognition by modelling it dynamically. As Chemero and Silberstein state: “Dynamical systems models are shown to work both in brain-only explanations and in brain–body-environment ones” (Chemero and Silberstein, 2008, p. 131).

In a similar tone, Van Gelder (1995) claims that the *dynamical hypothesis* is the unifying essence of dynamical approaches to cognition. It is encapsulated in the simple slogan, “cognitive agents are dynamical systems” (Van Gelder, 1995, p. 615). Van Gelder argues that the dynamic hypothesis has two aspects:

1. The first aspect is the nature hypothesis, which is about the nature of cognitive agents themselves, i.e. it considers that cognitive agents are dynamical systems. “Nature hypothesis is concerned in the first instance not with low-level systems but with how agents are causally organized at the highest level relevant to an explanation of cognitive performances, whatever that may be” (Van Gelder, 1995, p. 659).
2. The second aspect is the knowledge hypothesis. In cognitive science, cognitive agents can be understood dynamically, i.e. cognition can be understood in dynamical terms.

Randall Beer (1995) formalises a DST framework to evaluate the cognitive behaviour of an agent. It was applied to a six-legged agent, known as a hexapod. The framework that Beer developed was based on previous work on autonomous agents as well as work on the neural basis of animal behaviour. The agent is closely coupled with the environment, and their interactions are, in general, jointly responsible for the agent’s behaviour. Beer demonstrates the application of this framework by using it to synthesise and analyse the walking behaviour of a legged agent.

Beer describes an autonomous agent as an embodied system. It was designed to satisfy its internal and external goals through its actions, while being in a continuous long-term interaction with the environment in which it was situated (Figure 3.7). When modelling autonomous dynamical systems, the parameters are held fixed for the duration of any particular trajectory. Beer noticed that the central problem with an autonomous agent is how to generate appropriate behaviour at an appropriate time, because both its internal and external states change continuously. Here is an example: “An animal moving throughout its environment, needs to adopt many different modes of behavior as it becomes hungry or tired and encounters potential food, predators and mates, all the while adjusting its posture and leg movements to the constantly changing terrain which it is traversing” (Beer, 1995, p. 174).

An autonomous agent must be able to adapt flexibly to its immediate circumstances to meet its long-term goals. Thus, it must continuously adjust its behaviour in an appropriate way. Beer focuses on developing complete agents capable of carrying out open-ended tasks in an unconstrained environment, rather than agents with isolated cognitive skills in a restricted domain. A significant part of complex behaviour can emerge from the ongoing interaction between the agent and its environment.

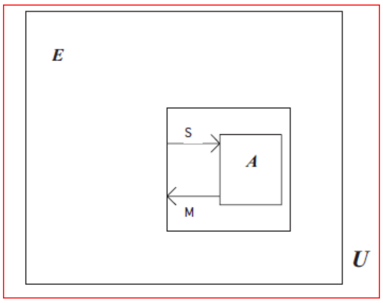


Figure 3.7. Organism–environment system U. Reproduced from Beer (1995, p. 182).

This figure shows the closely coupled system of an agent and its environment. In this figure, E is the environment and A is the agent. The agent and the environment together are one system, termed U (universe). *S* is the effect of the environment on or the input from the environment to the organism and *M* is the output of the agent or its action in the environment. To couple the two dynamical systems, the rate of change of the agent's state variables (x_{A}) depends on the environment's state variables (x_{E}) and vice versa. This coupling can be represented with a sensory function *S* of environmental state variables to give agent state variables and a motor function *M* from agent state variables to environmental state variables. Thus, $S(x_{\mathrm{A}})$ corresponds to the agent's sensory inputs, whereas $M(x_{\mathrm{E}})$ corresponds to its motor outputs.

The non-linear coupling of an agent and its environment can be modelled by the following non-linear equations:

$$\frac{dx_{\mathrm{A}}}{dt} = A\ (x_{\mathrm{A}}; S(x_{\mathrm{E}}); u_{\mathrm{A}})$$

$$\frac{dx_{\mathrm{E}}}{dt} = E(x_{\mathrm{E}}; M(x_{\mathrm{A}}); u_E)$$

where dx_{A} / dt is the instantaneous rate of change of the agent state variable x_{A} over time. It depends on the agent's state variable at the instant (x_{A}) and the impact of sensory inputs from the environment on the agent $S(x_{\mathrm{E}})$. Similarly, dx_{E} / dt is the instantaneous rate of change of the environment state variable x_{E} over time. It depends on the environment's state variable at the instant (x_{E}) and the impact of the agent on the environment $M(x_{\mathrm{A}})$.

The latter could be a motor action by the agent. Here u_A and u_E represent any remaining parameters for A and E, respectively, that do not participate in the coupling (Beer, 1995, p. 181).

It is assumed that this coupled agent–environment system exhibits only convergent dynamics. *S* is intended to represent *all* effects that E has on A. This influence occurs through what is normally thought of as a sensor. This breadth of usage is justified by the observation that any such effect can influence the subsequent trajectory of A. Likewise, *M* is intended to represent all effects that A has on E, whether or not they occur through what is normally thought of as an effector (Beer, 1995).

Any action that an agent takes affects its environment in some way through *M*, which in turn affects the agent itself through the feedback it receives from its environment via *S*. Likewise, the environment's effects on the agent through *S* are fed back through *M* to, in turn, affect the environment itself. Thus, each of these two dynamical systems is continuously deforming the flow of the other and therefore, influencing its subsequent trajectory. Note that one dynamical system cannot in general completely specify the trajectory of another dynamical system to which it is coupled (Beer, 1995).

3.3.3 DST and EC

Palermos (2014) uses Beer's theoretical framework for the coupling of two non-autonomous dynamical systems to establish that a feedback loop was a necessary and sufficient condition for the cognitive extension into the environment. This is Palermos's version of revised EM. The feedback loop is not only a sufficient condition, as Clark envisages, but a necessary condition for the extension of cognition into the environment. Palermos uses the revised version of EM to counter the criticism of Adams and Aizawa (A&A) regarding the coupling-constitution fallacy and cognitive bloat.

Palermos argues that Clark's formulation of the EM thesis failed to establish EC in a principled way. Therefore, the criticisms of C&C's EM thesis, such as the coupling-constitution fallacy and cognitive bloat, are relevant. Palermos identifies that the application of DST, especially feedback loops (i.e. continuous mutual interactions through non-linear relations between a cognitive agent and an external artefact) can address the coupling-constitution fallacy and cognitive bloat. Moreover, Palermos's aim was to clearly differentiate between HEC and HEMC.

Palermos argues that in an open-loop system, i.e. a one-way dependence system, the activity of the affected component has no ongoing direct effect on the affecting component. Thus, an external artefact causes only cognition and does not have a constitutional role in cognition.

To establish that a non-linear feedback loop is a necessary and sufficient condition, Palermos uses Beer's theoretical framework of autonomous and non-autonomous agents. Palermos considers an example of a cognitive agent (A) using a tactile–visual substitution system (TVSS). Palermos argues that the coupling of A and TVSS (ATVSS) is a perfect example of a dense non-linear relation between cognitive agent A and the TVSS. For these continuous-time non-autonomic dynamic systems, the dynamic laws will be *A* and *TVSS*, respectively. The agent and the substitution system are engaged in a continuous mutual interaction and the two systems are coupled non-autonomous dynamic systems. Because these systems are *coupled*, the rate of change of the *state variables* $u(t)$ of each system depends on the state variables of the other, and vice versa (Palermos, 2014a). In this mutual interaction, the rate of change of the agent's state variables depends on a function *E* of the substitution system's state variables. This function *E* captures all the ways in which the substitution system can affect the agent. Similarly, the rate of change of the TVSS's state variables depend on a function *I* of the agent's state variables, which encompasses all the possible ways in which the agent can affect the epistemic artefact. Thus, $E(x_{\mathrm{TVSS}})$ represents the effects of TVSS on the agent and $I(x_{\mathrm{A}})$ represents the effects of the agent on the TVSS. The coupled system is

$$\frac{dx_{\mathrm{A}}}{dt} = A(x_{\mathrm{A}}\ E(x_{\mathrm{TVSS}}), u_{\mathrm{A}}^{*})$$

$$\frac{dx_{\mathrm{TVSS}}}{dt} = x'_{\mathrm{TVSS}} = TVSS(I(x_{\mathrm{A}}), u_{\mathrm{TVSS}}^{*})$$

where u_{A}^{*} and u_{TVSS}^{*} represent any parameters in A and TVSS that are not affected by the coupling. Any action that the agent takes affects the visual substitution system in some way through *I*, which, in turn, affects the agent itself through the feedback it receives from the visual substitution system via *E*. Similarly, the visual substitution system's effects on the agent through *E* are fed back through *I* to, in turn, affect its own operation. Thus, each of the two dynamical systems is continuously deforming the flow of the other. We can think of these two coupled non-autonomous systems A and TVSS as a unified autonomous dynamical system, ATVSS (Palermos, 2014).

The coupling of two non-autonomous dynamical systems into an autonomous unified system can give rise to behaviours that goes beyond the sum of the behaviours that the individual subsystems can produce on their own. Thus, an agent's behaviour properly stems only from the dynamics of the coupled system (ATVSS) and not from the individual dynamics of either A or TVSS alone. The mutual interaction between the agent and its TVSS gives rise to new systemic properties that do not belong to either of the subsystems alone, but to the overall coupled system, ATVSS.

Palermos claims that for a system like ATVSS, we can formulate two distinct arguments for postulating that the subsystems are *coupled*:

(1) The properties of the coupled system cannot be attributed to any of the contributing systems alone, but to the *coupled* system as a whole. In other words, the coupling of the systems is necessary for accounting for the systemic properties, so they cannot be ontologically eliminated.

(2) For ongoing feedback loops between *coupled* systems, there is a dense non-linear causal interdependence that prevents us from decomposing the systems into distinct *inputs* and *outputs* from one to the other, since the effects of each component on the other are not entirely endogenous to the affecting component, and vice versa. Accordingly, we *cannot but* postulate that there is a coupled system. Overall, then, we might say that the *constituents* of the system are the interdependent components, which, because of feedback loops, give rise to the processes (and their properties) that we are interested in and which attracted our attention to the relevant components in the first place (Palermos, 2014, p. 33).

Having established the necessary and sufficient conditions for a coupled system via feedback loops, Palermos then reconsiders the examples of extended systems that the criticisers used for the coupling-constitution fallacy, cognitive bloat and the reduction to HEMC. Palermos argues that shopping lists and telephone directories are not examples of HEC as they do not meet the necessary and sufficient conditions required for HEC, that is, a non-linear, coupled feedback loop between the agent and the external artefact. In using a shopping list or a telephone directory, there is a causal one-way dependence rather than a feedback loop. Therefore, neither a shopping list nor a telephone directory is an example of HEC.

Palermos argues that concerns about cognitive bloat can be addressed by considering how an agent establishes a feedback loop with an external artefact:

> Ongoing mutual interdependence on the basis of feedback loops is the criterion by which we can judge whether two seemingly distinct systems constitute an overall system, consisting of both of them. Conversely, when no such mutual interaction is in play, but instead a system affects another one in an one-way dependence (i.e., the activity of the affected system has no ongoing direct effect on the affecting system, such that no feedback loops are exhibited), then we have a paradigmatic case of a merely causal—as opposed to constitutive—dependence. (Palermos, 2014a, p. 34)

Cognitive bloat can occur only with open-loop systems, but in HEC, open-loop systems are not examples of an extended cognitive system. In the same way, since a coupled feedback loop is essential for HEC, a system with a feedback loop cannot be described by HEMC, as its systemic properties cannot be individuated as being due either to the cognitive agent or to the external artefact. Therefore, there is a sharp distinction between HEC and HEMC.

Before Palermos, Chemero and Silberstein (2008) also asserted that a non-linear feedback loop was a necessary and sufficient condition for EM. Palermos agrees with Chemero but disagrees with the example of a feedback loop that Chemero presents, i.e. the non-linear relation between an outfielder and a fly ball. Palermos argues that Chemero has an incorrect understanding of non-linear relations. Palermos states that: "These *non-linear relations arise only out of cooperative or inhibitory feedback loops between interacting parts*" (Palermos, 2014, p. 36). Contrary to what Chemero claims, the objects of perception are neither non-linearly related nor, thereby, coupled to their perceiver; there are no feedback loops in perception. Instead, the agent is only linearly dependent on the objects it perceives. Palermos concludes that Chemero's misunderstanding of the nature of non-linear relations seems to lead him back to the fallacy he has previously offered a solution for (Palermos, 2014).

3.3.4 Limitations of Palermos's Arguments

Palermos applies DST to cognition as in the DST modelling of an agent with TVSS to address the criticisms of EM, such as the coupling-constitution fallacy and cognitive bloat. Palermos claims that the dense, non-linear close coupling of an external device, such as a TVSS, with the agent via feedback loops is an example of HEC that is immune to the coupling-constitution fallacy and cognitive bloat. Palermos also challenges Rupert's claim that, by using DST, HEC will dissolve into HEMC, such as for an agent and TVSS where the cognitive success cannot be individuated to the agent or to the TVSS. The agent and TVSS are a constitutive whole in the accomplishment of the cognitive success. Therefore, there is a sharp distinction between HEC and HEMC. It is important to note that both the application of DST to cognition and its claim that such dynamic modelling of cognition ensures HEC are not free from criticisms.

The core of the criticism is that DST explains intelligent behaviour but it fails to explain how that behaviour originated. Clark argues that: "Commanding a good pure dynamical characterisation of the system falls too far short of possessing a recipe for building a system that would exhibit the behaviours concerned" (1997a, p. 120). Proponents of DST claim that applying DST to the mind implies that cognition extends into the environment due to the non-linear coupling of the environment with the agent. However, Shapiro (2013) argues that applying DST to psychology does not guarantee the extension of mental states that the proponents of DST normally claim.

Clark states that the DST account of cognition is mainly descriptive. For example, Thelen (1995) describes an experiment demonstrating the A-not-B error. In an A-not-B task, a child witnesses a toy being hidden under box A and will successfully retrieve it from there. However, after witnessing the toy being moved and hidden under box B, the child will continue to search under box A. This mistake is the A-not-B error. Thelen's DST model describes the interplay between environmental settings, the cuing paradigm (how the child is prompted and signalled) and the child's past reaching behaviours to explain why the A-not-B error occurs. However, Clark points out a limitation that although DST can aptly describe behaviours and their various influences, Thelen et al. do not, however, identify the underlying mechanisms that are responsible for the A-not-B error.

For Beer's DST model of a legged agent, Clark states that:

> This kind of geometric, state-space-based understanding is, to be sure, both valuable and informative. It remains an open question, however, to what extent such explanations can replace, rather than merely complement, more traditional understandings couched in terms of computational transitions and inner representational states. (Clark, 1997a, p. 120)

Clark argues that the radical position of DST, which advocates the wholesale replacement of computation and representation by geometric dynamical systems, faces two crucial challenges:

1. Scaling and tractability: Even a 30-neuron leg controller constitutes a dynamical system of such complexity that our intuitive geometric understanding breaks down. Moreover, the detailed mathematics of DST becomes steadily less tractable as the number of parameters and the size of the state space increase. As a result, Beer's analysis was, in fact, conducted only for a simpler five-neuron system controlling a single leg. The practical applicability of DST to highly complex, high-dimensional, coupled systems (like the human brain) must, therefore, be in serious doubt.

Beer's latest work on *Caenorhabditis elegans* (*C. elegans*) a transparent worm that lives in temperate soil environments, reaffirms the issues raised by Clark. It was the first animal to have its genome fully sequenced (Izquierdo & Beer, 2016). The behaviour of *C. elegans* is driven by 95 body wall muscles operating on the nematode's hydrostatic skeleton. This musculature is activated by a total of 302 neurons. Although there have been many breakthroughs in dynamically modelling the whole worm (brain, body and environment), there are still many challenges, such as:

> A better understanding of force development in the body wall muscles and its interaction with the hydrostatic skeleton and surrounding medium, and the incorporation of neuromodulatory effects into electrophysiological models. In addition, open source integrative efforts such as the Open Worm initiative have an essential role to play in coordinating and supporting multiple modeling efforts. (Izquierdo & Beer, 2016, p. 27)

Even for a 1-mm *C. elegans*, DST modelling is too complex and difficult for humans.

2. The second and more fundamental challenge concerns the type of understanding such analyses provide. For example, the DST model of Kelso's work on the wagging behaviour of fingers can be used to predict the movement of fingers, but such an analysis does not provide a full explanation of how such movement arises. Clark expresses his concerns over the type of understanding that can be provided by a dynamical analysis. He argues that a dynamic analysis provides an abstract explanation rather than a full explanation:

> This type of understanding threatens to constitute an abstract description rather than a full explanation. We learn what the system does and when it does

> it, and what patterns of temporal evolution its behaviour displays, but this understanding, although valuable, does not seem to be exhaustive. In particular, we are often left – as I will later argue in detail – with an impoverished understanding of the adaptive role of components, and of the internal functional organisation of the system. (Clark, 1997, p. 101)

It is generally considered that cognition is the cause of behaviour; however, the proponents of DST consider "cognition as intelligent behaviour" (Chemero, 2009). Shapiro rejects this conception of cognition and behaviour in DST, as it results in either confusion or brazenness:

> Confusion, in as much as it equates causes and effects: cognition cannot be both cause of behaviour *and* behaviour. Or brazenness, to the extent that there seems no good reason to insist that cognitive science abandon its goal of describing the causes of behaviour in terms of representational capacities.
>
> (Shapiro, 2013).

Shapiro argues that the explanandum of DST is to study intelligent behaviour rather than cognition itself. It is one thing to use equations to describe how a system behaves and quite another to uncover the causes of this behaviour. Shapiro (2013) argues that the EC thesis does not follow from the conclusions of dynamical cognitive science unless one is prepared to accept cognition as intelligent behaviour. Shapiro also claims that one can embrace dynamical cognitive science without also believing that cognition can be extended or that representation must be eliminated from explanations of cognition.

From the criticisms of Clark and Shapiro, it can be concluded that although DST is effective in predicting the behaviour of a cognitive agent in a geometrical space, it does not provide much detail about how cognition arises in the first place. DST can be applied to a neural network; however, as shown for *C. elegans*, a multi-neuronal network is too complex to model. Shapiro's argument is aligned with Rupert's argument that DST can apply to HEMC to explain the behaviour of a cognitive agent. If that is the case, how can DST unambiguously support HEC?

Rupert (2009) argues that, in the application of Beer's DST framework to organismically internal dynamical interactions, such as those in hexapod locomotion, "these are purely internal models, however, and as such do not support the extended view" (Rupert, 2009, Chapter 7, p. 12). Rupert claims that Beer's DST framework suggests that a hexapod is an embodied agent, which aligns with the traditional boundary between an organism and its environment. It does not support the extended view. Rupert (2009) “argued that the dynamical-systems-based models that most clearly apply to cognitive tasks also mark a traditional boundary between the organism and the environment, a boundary mediated by sensory (or sensory-like) inputs” (Rupert, 2009, Chapter 7, p. 1).

To get the bottom of this criticisms, I will critically assess Beer’s theoretical framework and its application to a legged agent as well as Palermos’s example of TVSS with respect to the criticism raised by Rupert. Figure 3.8 shows the hierarchical control architecture for a six-legged agent.

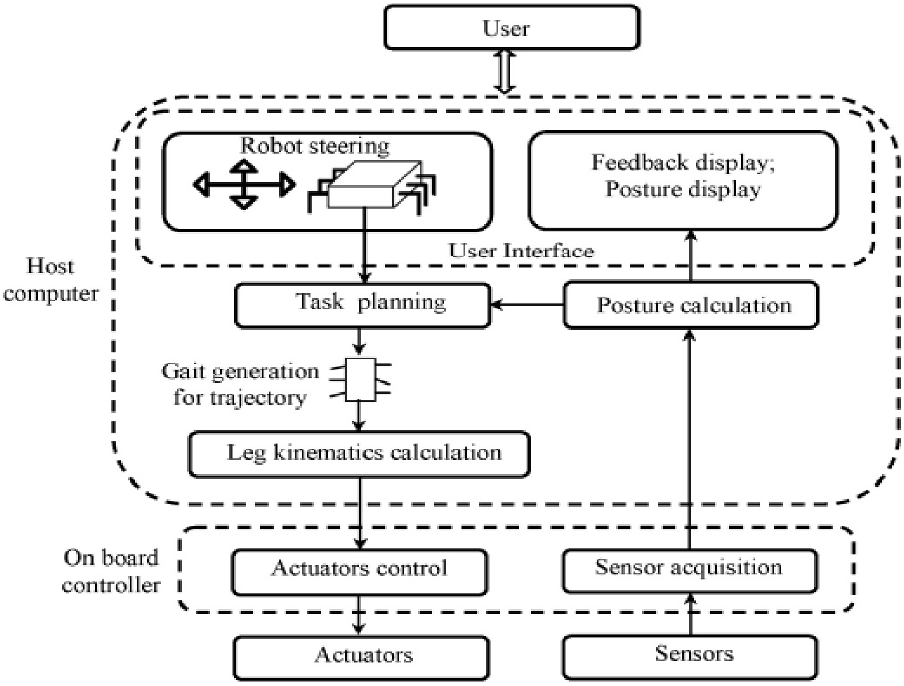


Figure 3.8. Control architecture for a six-legged agent. Reproduced from Tedeschi and Carbone (2014, p. 196).

> The six-legged agent comprises an artificial neural network that allows for the simulation of a considerable amount of behavioural data. A number of properties observed emerge from a decentralised architecture. Examples are the continuum of so-called gaits, coordination of up to 18 leg joints during stance when walking forward or backward over uneven surfaces and negotiation of curves, dealing with leg loss, as well as being able following motion trajectories without explicit pre-calculation.
>
> (Schilling et al., 2013, p. 397)

As can be seen, the hexapodal robot is like an embodied agent with sensors that uses its body for locomotion. We can say that it is part of a closely coupled system with a brain (neural network), body and environment.

Now let us consider Palermos's example of an agent and a TVSS (Figure 3.9). A typical sensory substitution system has three major components: (a) a sensor that senses information that would have been received by the substituted modality, typically vision, (b) a coupling system that processes the sensor's output and drives an actuator and (c) an actuator that activates the receptors of the substituting modality, such as skin mechanoreceptors or auditory hair cells (Bach-y-Rita & Kercel, 2003). In a TVSS, a camera is used to detect an object and the signal is converted to an electrical signal, which is passed to a sensitive skin surface, such as a fingertip or the tongue. With active sensing, motor–sensory relations, and not sensory signals per se, are the relevant cues for the perception of external objects.

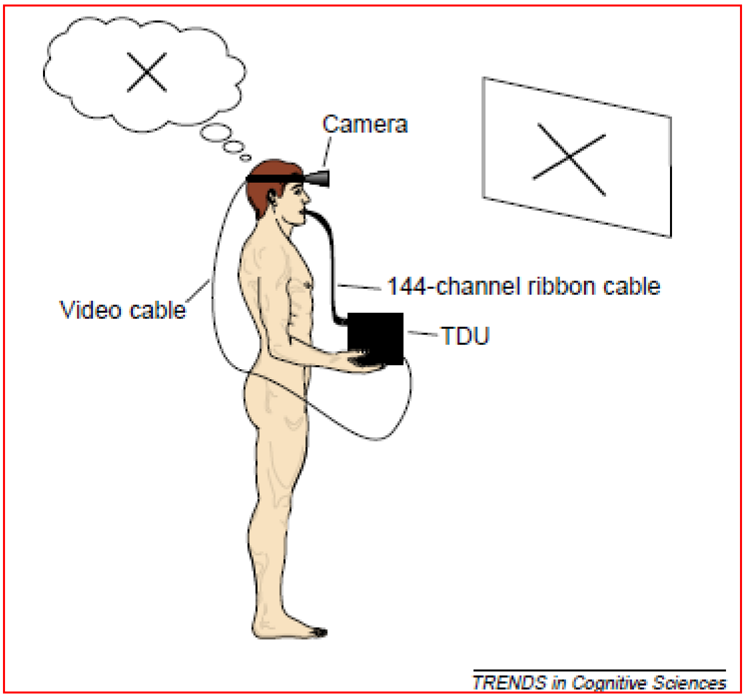


Figure 3.9. Schematic of a tactile-vision substitution system (TVSS). Reproduced from Bach-y-Rita and Kercel (2003, p. 543).

In the system shown in Figure 3.9, an image is captured by a head-mounted CCD camera. The video data are transmitted to the tongue display unit via a video cable. The tongue display unit converts the video signal into a pattern of 144 low-voltage pulse trains, each corresponding to a pixel. The pulse trains are carried via a ribbon cable to a flexible electrode array placed on the dorsum of the tongue in the mouth. The electrodes stimulate touch sensors on the dorsum of the tongue as electrotactile stimuli. The subject can experience the resulting stream of sensations as an image (Bach-y-Rita & Kercel, 2003, p. 543). After training with TVSS, subjects report experiencing images in space, instead of on the skin. They learn to make perceptual judgements using visual means of interpretation, such as perspective, parallax, looming and zooming, and depth estimates (Bach-y-Rita and Kercel, 2003, p. 543). No manipulation of the TVSS is required by the agent to achieve the perceptual judgement.

Sensory substitution relies on brain plasticity, which, as the name indicates, is due to the plastic and flexible nature of neurons. The slogan for brain plasticity in neuroscience is: "Cells that fire together wire together, whereas cells that fire apart will wire apart." The neurons that fire together, e.g. during visualisation, are the visual, auditory and tactile neurons in the occipital cortex. Together, they achieve a visual experience. However, if one of type of sensory neurons, say the visual sensory neurons, are impaired, then other auditory and tactile sensory neurons, which normally fire with the visual sensory neurons, will be flexible enough to generate a kind of visual experience from touch and hearing. This is brain plasticity. It is fundamental for sensory substitution. In short, sensory substitution allows someone to perceive environmental information that is normally received via one sense (like vision) via another sense (like touch or hearing).

Both the hexapod and an agent with a TVSS do utilise sensors to establish feedback loops with their environment. This is evident in how the hexapod navigates its environment and in how the TVSS translates visual data into tactile sensations for the agent. Based on Rupert's criticism, one might argue that Beer's DST applied to the hexapod would suggest that embodied cognition occurs such that the hexapod's legs and structure play a significant role in its cognitive processes. The interactions between the hexapod's body and its environment contribute to its behaviour. This does not mean that its cognition extends into the environment; instead, it highlights how the body's structure and dynamics can shape cognitive processes. In the same way one might argue that the agent's interaction with a TVSS might be seen as an enhanced sensory process rather than a cognitive one, i.e. the TVSS is merely translating one form of sensory information into another. Accordingly, the interaction between the agent and the TVSS is not necessarily a form of EC but rather shows how artefacts can augment sensory capabilities. The underlying cognitive processes could still be occurring internally within the agent. Although the TVSS and hexapod examples provide interesting contexts for exploring cognition, Rupert emphasises the traditional boundaries of cognitive systems and the importance of embodied and embedded rather than extended cognition. Rupert could argue that although the TVSS is causally coupled to the agent, i.e. it affects and is affected by the agent, that does not mean that the TVSS plays a constitutive role in the agent's cognitive processes. According to Rupert's criticism explained earlier, although DST can describe feedback loops and the mutual interdependence between a system and its environment, that does not inherently advocate for the EC view.

Palermos, however, can argue that there is a continuous ongoing feedback loop between the agent and the TVSS, as follows. The TVSS provides the agent with sensory feedback. The agent adjusts their behaviour based on this feedback, and the TVSS adjusts its feedback based on the agent's responses. There is a continuous loop. When an agent uses a TVSS, they are reliant on it to interpret visual information from their environment, which they perceive tactilely. In turn, the TVSS relies on the agent's tactile sensitivity and their ability to interpret the tactile stimuli as visual information. Thus, there is a tight feedback loop in which both systems are mutually dependent on each other. Ongoing feedback loops ensure that the interaction between the agent and the TVSS is dynamic and reciprocal. This continuous interplay, in which both the agent's actions influence the TVSS and the TVSS's state influences the agent, forms an integrated system, as opposed to a mere one-way interaction. This interdependence mirrors the DST framework, which suggests that such systems can be considered as a single dynamic system. Using feedback loops as a defining criterion means that the agent and the TVSS, when interconnected, form a single cognitive system. The agent's perception and cognitive processing are no longer just a product of internal mechanisms but actively incorporate the external TVSS. The agent's cognitive processes have effectively extended to incorporate the TVSS, not just as a tool, but as an integral part of their cognitive architecture. Therefore, Palermos can conclude that the example of the agent and TVSS shows that when two systems exhibit an ongoing mutual interdependence through feedback loops, as outlined in Beer's DST framework, they can be considered as parts of an extended cognitive system. This blurs the boundaries between an individual and their environment in terms of cognitive processing.

Favela et al. (2021) discuss how their research supports EC, particularly in the context of sensory substitution devices being used like a rod or an enactive torch in affordance judgement tasks. The study focuses on how people make perceptual judgements about their ability to perform actions in an environment using tools. It demonstrates that the cognitive dynamics involved in these judgements extends beyond the individual's body to include the tools they use. The research by Favela et al. (2021) provides a compelling empirical foundation for understanding EC through the lens of DST, particularly in the context of sensory substitution. Their findings highlight how cognitive processes can extend beyond the brain and body to include external tools, thus forming a dynamic, adaptive system. This integration, characterised by feedback loops and non-linear interactions, is not merely an enhancement of sensory capabilities but a fundamental transformation of cognitive processing and affirms the principles of EC and DST in practical scenarios.

Although Palermos's arguments provide valuable insights by suggesting that the ongoing feedback loop between an agent and an artefact is a necessary and sufficient condition for EC, I would like to see whether there is a somewhat more nuanced way of understanding the necessary and sufficient condition for cognitive extension, especially by considering Rupert's criticism.

Consider the previous example of a cognitive agent using NVGs (Pritchard, 2018a). When an agent wears NVGs, the device does not merely provide information, since it alters the agent's perceptual experience. Over time, the agent perceives the environment directly through the goggles, rather than interpreting it being as mediated by a tool. The agent's movements and focus, such as where to look, influence the input from the NVGs. Simultaneously, what the NVGs display informs the agent's subsequent actions. This continuous feedback loop between the agent's decisions and the NVGs' input and output resembles the dynamic interaction critical for EC. As the agent becomes accustomed to the NVGs, the boundaries between their natural vision and the enhanced vision from the goggles might start to blur. They might operate as naturally with the goggles as without, indicating that there has been a deep integration of the device into their cognitive processes. Based on the above, the use of NVGs by an agent is not just a mere enhancement of their vision but is an example of EC, as the device has become an integrated part of the agent's perceptual and cognitive system. The agent's perception of their entire environment can be transformed through the NVGs, which encompass everything within their visual field. This comprehensive transformation means that the agent's entire perceptual cognitive environment is mediated through the goggles, lending more weight to the actuality of EC.

Carl Craver's (2007) mutual manipulability (MM) provides insights about the manipulation of artefacts that are relevant to our discussion of DST-based EC. Craver

(2007) outlines the principles of MM as part of his broader discussion on mechanistic explanations in neuroscience. He argues that understanding a mechanism involves demonstrating how manipulating one part of a system causally affects other parts and vice versa. This approach can help us to distinguish between components that are merely causally related to a system and those that are constitutively part of it, and it provides a broader context for understanding Craver's contributions to the philosophy of science, particularly regarding how components of a system are understood to be causally and constitutively related within mechanistic explanations. MM is a significant tool for understanding how components contribute to a system, particularly in the context of cognitive systems. MM revolves around the idea that if manipulating a component X leads to changes in a system S and that manipulating S leads to changes in X, then X can be considered a constitutive part of S. This criterion is used to establish a non-causal, constitutive relationship between parts of a mechanism or system. In other words, it helps us to identify parts that are essential to the functioning of the whole. MM helps to differentiate between components that are merely causing effects in a system (causal relationship) and those that are integral to the system's operation (constitutive relationship). Now let us apply MM to the agent and NVGs (ANVG):

Initial Stage of the Relationship Between the Agent and the NVGs: Initially, the soldier views the goggles as an external instrument. She is hesitant to rely on them fully when forming beliefs due to unfamiliarity. This stage is characterised by a causal relationship in which the goggles merely assist in perception without being an integrated part of the cognitive process.

Developing Proficiency and Trust: As the soldier becomes more adept at using the goggles, her interaction with them changes. She begins to form beliefs based on what she sees through the goggles. This transition marks the beginning of a deeper, more integrated relationship between the soldier and the device.

Applying MM to Manipulating the Goggles (Tool): If altering how the goggles function (such as adjusting settings or changing the display mode) affects the soldier's cognitive processes (like decision-making, attention or belief formation), then this suggests that the goggles are becoming a constitutive part of these processes.

Manipulating Cognitive Processes: Conversely, if changes in the soldier's cognitive strategies (like where she decides to look or what she chooses to focus on) lead to changes in how she uses the goggles, this supports their role as being constitutive of her cognitive processes.

Integration into Cognitive Processes: The soldier's increasing reliance on the goggles for perception and for the formation of beliefs indicates that the goggles are no longer just tools. They have become an integrated part of her cognitive system, altering and being altered by her cognitive states and processes.

Feedback Loops and Cognitive Extension: The development of feedback loops – in which what the soldier sees influences her subsequent actions and adjustments of the goggles – further exemplifies MM. The goggles are not only influencing cognitive processes but are also being influenced by them, which demonstrates the reciprocal relationship.

Conclusion – Goggles as Being Constitutive of Cognition: In this example, the NVGs transition from being an external instrument to an integrated part of the soldier's cognitive system. This integration aligns with EC, as cognition is not confined to the brain but extends to tools actively manipulated and integrated into cognitive processes.

The lens of MM shows that the NVGs are constitutive of cognitive processes, as changes in their use and functionality directly affect and are affected by the soldier's cognitive states and actions. Applying Craver's MM to this example demonstrates how, over time, the NVGs become an integrated part of the soldier's cognitive system, which supports EC. This integration is evident in the reciprocal influence between the use of the goggles and the soldier's cognitive processes and illustrates that there is a constitutive relationship beyond mere causal interaction.

Farina and Lavazza (2022) highlight the distinction between passive and active tool use. Active engagement with a tool, especially devices that interact closely with cognitive processes, can lead to a form of integration in which the tool is no longer an external aid but a constituent part of the cognitive process. This supports a DST perspective of extended cognition, one in which the system (the cognitive process in this case) is not static but adapts and evolves with new elements, as for ANVG. DST models how these systems can dynamically integrate new components and is aligned with Farina and Lavazza's views on active tool usage. ANVG is a concrete example of how the active manipulation of a device leads to cognitive modifications. Farina and Lavazza (2022) discuss how the proficient use of sensory substitution devices results in a reconfiguration of neural processes and perceptual experiences, effectively extending the cognitive system to include the device. This empirical example illustrates key principles of DST in the context of EC. It shows how complex systems (like the human cognitive system) can incorporate new elements and reorganise their functioning, thus leading to emergent cognitive capabilities.

Although both the examples considered – ATVSS and ANVG – involve feedback loops, the soldier's use of NVGs requires active manipulation. The act of manipulation involves a combination of perception, intention and motor action, thereby enriching the overall cognitive process. This active engagement might be seen as a clearer demonstration of how external artefacts can become an integrated part of our cognitive processes. The soldier physically adjusts, focuses and directs the goggles based on the situation. This direct manipulation and control over the artefact highlights that there is a deeper integration of the tool into the cognitive processes. The act of manipulating the NVGs – adjusting the focus, direction and even the mode of vision – amplifies the feedback loop. It is not just a static feedback loop, as the loop is continuously shaped and refined by the agent's actions, which strengthens the case for EC. The agent has the autonomy to choose when and how to use the NVGs, whereas the TVSS provides feedback somewhat passively. The agent's intentional actions with the NVGs – choosing when to use them, how to adjust them and where to look – bring a level of agency to the interaction. It is not just about receiving feedback; it is about actively shaping the feedback process. Although both examples – the agent and NVGs and the agent and a TVSS – are strong cases for EC, Pritchard's example of the agent and NVGs offers a more compelling case due to the direct and active manipulation of the external artefact by the agent. This manipulation, combined with the feedback loop, underlines that there is a deeper and more integrated cognitive extension. As the agent gains experience in using the NVGs, their proficiency improves, leading to a refined extended cognitive process. This evolution, driven by both feedback and active manipulation, illustrates that there is a deepening of the cognitive extension over time. Based on the above, Pritchard's example of the agent actively manipulating NVGs, in conjunction with the feedback loop, provides a dynamic and deeply integrated demonstration of EC, making it a more comprehensive example of EC. The NVGs scenario is a more direct and active exploration of how external artefacts can become deeply integrated into our cognitive processes. The immediacy of the consequences, the active manipulation and the enhancement of a natural sense arguably provide the NVGs scenario with more explanatory potential. Based on the above, by emphasising the dual conditions of manipulation and feedback, one can argue for a richer, more dynamic and interactive understanding of EC, which might capture the complexity of human–tool interactions more comprehensively than relying on feedback loops alone. By combining manipulation with feedback, this approach offers a more robust framework for understanding the intricate interplay between agents and their artefacts, making it a more explanatory and powerful model than relying on feedback loops alone.

Although, both ATVSS and ANVG capture essential aspects of EC, they have different focal points. Palermos's model underlines the essentiality of feedback and the continuous interplay between the agent and the external system. This approach captures the fundamental interconnectedness. Pritchard's model, by highlighting both feedback and manipulation, provides a more dynamic view of cognitive extension. It recognises the agent's active role in shaping the interaction, making it more aligned with many real-world scenarios in which humans actively engage with artefacts.

Based on the above arguments, I conclude that the manipulation of an external artefact and existence of a feedback loop are necessary and sufficient conditions for cognitive extension and that this has better explanatory power than a feedback loop alone. From an explanatory perspective, the ANVG scenario offers a more comprehensive framework for exploring EC. Its emphasis on active tool manipulation, combined with intricate feedback loops and the multifaceted nature of cognitive tasks, makes it a robust model. Its emphasis on active manipulation, deep integration of the artefact and complex feedback loops makes it a more robust exemplar of the principles of DST and EC.

Let us return to Rupert's preference for HEMC over HEC. Although Rupert's embedded cognition provides valuable insights into the relationship between cognition and the environment, EC, based on our discussions, seems to offer a more nuanced, dynamic and holistic understanding of how external artefacts can become integrated into cognitive processes.

Although Rupert's argument for simplicity in favour of HEMC is compelling, the result may be to oversimplify complex cognitive phenomena. In cognitive science, the most parsimonious explanation is not always the most accurate or comprehensive, especially for complex, dynamic systems. EC provides a more nuanced understanding of the interplay between cognisers and their environments as it recognises the deep integration of external artefacts in cognitive processes. This comprehensive approach may be necessary to fully explain the intricacies of how cognition extends beyond the brain. HEMC suggests that external tools and the environment play an important role in supporting cognitive processes but are not actually part of the cognitive system itself. Although HEMC explains how external artefacts can aid or influence cognition, it may fall short in explaining scenarios in which these artefacts are deeply integrated into the cognitive process. In cases where external artefacts and the cogniser are engaged in a dynamic, reciprocal relationship, viewing these artefacts as merely embedded tools fails to capture the full extent of their role in cognition. HEMC may not fully account for the degree to which cognitive processes are altered, shaped or even constituted by these artefacts. The major difference between EC and HEMC is the importance of dynamic interactions and reciprocal causation. If an external artefact and a cogniser are engaged in a continuous feedback loop, each is actively influencing the other. This mutual influence suggests a level of integration that goes beyond the artefact simply being an embedded tool. The cognitive process evolves and adapts in response to this interaction, thus indicating that the artefact plays a constitutive role in cognition.

Farina and Lavazza (2022) address Rupert’s criticism of EC by differentiating between the passive and the active use of tools and by emphasising how active tool use can lead to cognitive modifications that would not be possible by the brain alone. Rupert’s critique primarily focuses on the idea that the extension of cognition into the environment (via tools or other external resources) does not necessarily alter the fundamental nature of cognitive processes. Farina and Lavazza (2022) suggest that although Rupert’s critique may apply to passive tool use, it does not hold for active tool usage, which involves tools that interact with the user’s cognitive processes in a way that fundamentally alters or enhances them.

The following discussions, i.e. the argument from developmental systems theory, the argument from niche construction and cognitive niche construction theories, and the argument from the patterns in external information-bearing structures, show that feedback loops between an agent and artefact and the manipulation of the external artefact by the agent are fundamental to cognition.

3.4 Argument from Developmental Systems Theory

In this section, I focus on the systems approach, the developmental systems approach and the application of DST to developmental systems theory. Both EC and developmental systems approaches cater for the complex, emergent and interconnected nature of systems by accounting for the feedback loop between an agent and the environment. A feedback loop between the agent and their environment is fundamental in shaping cognitive and developmental processes, such as adaptation, learning and the emergence of new behaviours or properties, making it central to both EC and the developmental systems approach.

Unlike analytical methods that focus on constituent parts, system theories adopt a holistic and integrated approach. Current developments in the study of knowledge are like watertight compartments, as they focus on piecemeal and in-depth investigations. Although such methodologies have helped to achieve great leaps in the sciences and social sciences, they fail to provide an integrated world outlook. The slogan for systems theory is that the whole is more than the sum of the parts. The systems approach is an integrated approach.

Developmental systems theory focuses on how a single cell becomes a full-fledged organism. For example, a zygote, the union of a sperm cell and an egg cell, grows into an infant, which becomes an adult. One of the core aspects of development is that new patterns emerge from existing patterns. Developmentalists look for what causes an emergent new pattern in the inbuilt genetic code and the environment in which the organism grows. Biological systems are open systems and they continuously exchange matter, energy and information with the environment. A living organism absorbs energy from the environment and grows. As a result, new patterns emerge by self-organisation, that is, "pattern *and order emerge from the interactions of the components of a complex system without explicit instructions*, either in the organism itself or from the environment. Self-organization – processes that by their own activities change themselves – is a fundamental property of living things" (Thelen & Smith, 2006, p. 259).

There is a strong interrelation between nature and nurture during development. Here, a feedback loop with the environment is an inseparable part of the development of an organism. DST can apply to a developmental system. As Thelen and Smith (2006) argue, the application of DST to developmental systems results in two themes:

1. Development can be understood only as the multiple, mutual and continuous interactions among all the levels of the developing system, from the molecular to the cultural.

2. Development can be understood only as a series of nested processes that unfold over many timescales, from milliseconds to years (Thelen & Smith, 2006 p. 258).

Developmental systems theory describes feedback loops and changes over time (Figure 3.10). Ford and Lerner propose that humans rely on the “multilevel, contextual organisation of structures and functions” with varying kinds of stability and variability. This organisation of structures and functions can change both in and between levels. Thus, the theories behind biological and social development can be combined into a single developmental systems theory (Thelen & Smith, 2006 p. 270). Developmental theories also explain the dynamic nature of development, which is aligned with niche construction theory’s (NCT’s) ontogenetic and cultural feedback loop between an organism and its environment, which I discuss later.

Figure 3.10 shows Ford and Learner’s model of developmental change as a series of probabilistic states, where control systems interact in the person and the environment. States are thus the current configuration of the system, based both on current status and on the system’s immediate and long-term history.

According to these theorists:

> The definition, they maintain, implies a lifelong possibility of change, multiple (although not infinite) and nonlinear developmental pathways, discontinuities, and the emergence of new forms. Furthermore, the definition specifies that development is never a function of person or context alone, but indeed results as a function of their dynamic interaction.
>
> (Thelen & Smith, 2006, p. 270)

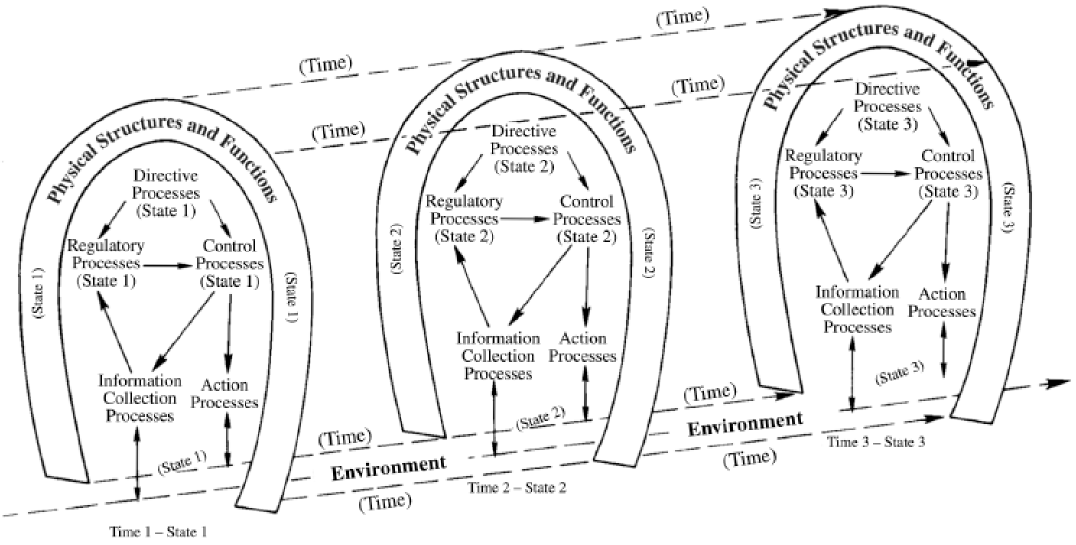

(Time)
(Time)
(Time)
(Time)
(Time)
(Time)
(Time)
Physical Structures and Functions
Directive Processes (State 1)
Regulatory Processes (State 1)
Control Processes (State 1)
Information Collection Processes
Action Processes
(State 1)
(State 1)
(State 1)
Time 1 – State 1
Directive Processes (State 2)
Regulatory Processes (State 2)
Control Processes (State 2)
(State 2)
(State 2)
(State 2)
Time 2 – State 2
Directive Processes (State 3)
Regulatory Processes (State 3)
Control Processes (State 3)
(State 3)
(State 3)
(State 3)
Time 3 – State 3
Environment
Environment

Figure 6.5 Ford and Lerner's model of developmental change as a series of probabilistic states.

Figure 3.10. Dynamic cognition. Reproduced from Thelen and Smith (2006, Chapter 6, p. 270).

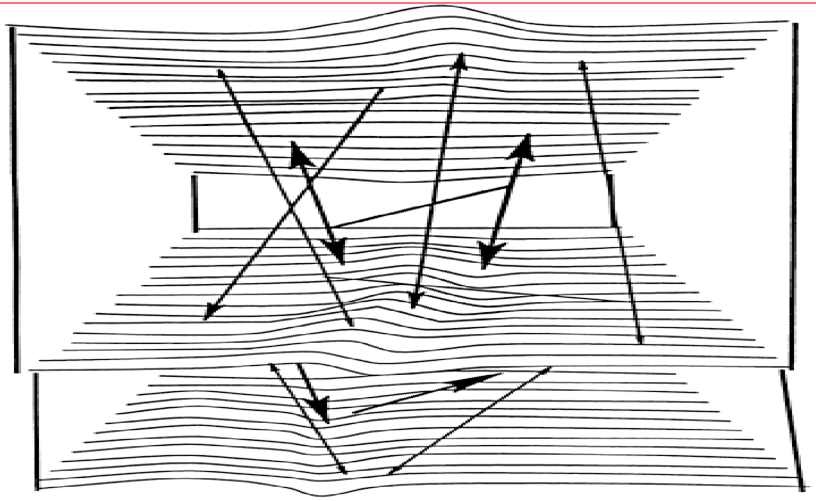


Figure 6.10 The epigenetic landscape as a multilayered system where the components mutually influence each other in changing ways.

Figure 3.11. Epigenetic development landscape. Reproduced from Thelen & Smith (2006, Chapter 6, p. 281).

Figure 3.11 shows an epigenetic development landscape as a multi-layered system.

> That depiction shows three landscapes layered on top of one another, indicating that the components of the dynamic system themselves have a dynamic. The arrows connecting the layers show that the coupling between the components is complex and contingent, and may change over time. This

> means that the coupling is always multidirectional, and that effects of the subsystems on one another may cascade over time.
>
> (Thelen & Smith, 2006, p. 280)

3.5 Argument from NCT

In this section, I evaluate how NCT and cognitive niche construction theory can support the role of feedback loops between an agent and external information-bearing structures in cognition. Section 3.5.1 focuses on the background of NCT. Section 3.5.2 differentiates NCT from standard evolution theory (SET), while Section 3.5.3 describes human niche construction. Section 3.5.5 considers social niche construction, and Section 3.5.6 analyses cognitive niche construction theory and cognitive niche theory. Both EC and NCT describe the dynamic, reciprocal interaction between an agent and their environment or an artefact. The agent modifies and interacts with the environment or tool, which in turn influences the agent. This interaction is not one way; it is a continuous feedback loop in which both the agent and the environment/tool are influencing and being influenced by each other.

3.5.1 Background

As Wilson and Clark (2009) state, developmental systems theory and the biological explanation of niche construction emphasise the need for cognitive activities beyond the skin and the skull, which support the claim for an EM. In biology, a niche is sometimes defined as the role an organism occupies in an ecosystem. Niche construction is the legacy of Lewontin: "The organism influences its own evolution, by being both the object of natural selection and the creator of the conditions of that selection" (Levins & Lewontin, 1985, p. 106). An environmental modification by an organism is called niche construction. NCT has two parts: (1) environmental modification by an organism (niche construction) and (2) the role of ecological inheritance, i.e. the legacy of the environmental changes made by the organism over time in the evolutionary process.

Gupta et al. (2017, p. 492) state that in SET, i.e. neo-Darwinian evolutionary theory, which unified Mendelian genetics with Darwin's theory of natural and sexual selection, the relationship between an organism and its environment can be expressed as two asymmetrically coupled differential equations:

$$\frac{dO}{dt} = f(O, E) \qquad (1)$$

and

$$\frac{dE}{dt} = g(E) \tag{2}$$

where O represents the organism and E represents the environment. The first equation shows that an infinitesimal change in the organism with respect to time is a function of the organism (*O*) and its environment (*E*). The second equation shows that an infinitesimal change in the environment over time is a function only of the environment (*E*). As per Lewontin (1983, 2000), the second equation does not fully capture the change in the environment resulting from the activities of the organism. Therefore, Lewontin changes the second equation to

$$\frac{dE}{dt} = g(O, E) \tag{3}$$

This third equation recognises an infinitesimal change of the environment over time as a function of the organism (*O*) and its environment (*E*). At the core of NCT are the environmental modifications by the organism and its ecological inheritance, i.e. the changes to the environment by the organism and the resulting ecological inheritance, which acts like an evolutionary process. NCT captures the Lewontin theme of the triple helix approach to evolution based on the co-equal and interacting dimensions of genes, the organism and its environment (Levins & Lewontin, 1985).

However, Gupta et al. (2017) argue that, although Equation (3) indicates that there is a superficial symmetrical relation in the function *g* (*O*, *E*), the consequences of *O* and *E* can differ. "The weightage of the biological phenomena referenced by these two functions will typically be different, as will their consequences for the joint dynamics of *O* and *E*, even though the two functions look superficially symmetrical in form" (Gupta et al., 2017, p. 492).

3.5.2 NCT versus SET

In SET, organisms are not treated as the cause of any evolutionarily significant changes in the natural selection pressures in their environments. In SET, natural selection pressures are solely due to genetic inheritance. However, in NCT, the activities of organisms are treated as co-causes of evolutionarily significant changes in the environment. In NCT, the evolution of an organism is assumed to be directed by two reciprocal processes:

1. Natural selection: The transmission of genes from parents to their offspring through genetic inheritance is influenced by natural selection.
2. Niche construction: In niche construction, selected habitats, modified and "constructed" habitats, and modified sources of natural selection in those habitats are also transmitted by a niche-constructing organism to its descendants by a second general inheritance system called ecological inheritance (Odling-Smee & Turner, 2011, p. 285).

Figure 3.12 shows the differences between SET and NCT. In SET [Figure 3.12(a)], niche construction is considered to be a product of natural selection. However, in NCT [Figure 3.12(b)], niche construction is recognised as an evolutionary process in which ecological inheritance has a parallel role to genetic inheritance. An organism and its environment have a reciprocal causal relation in which niche construction is a cause of evolutionary change rather than an effect of SET.

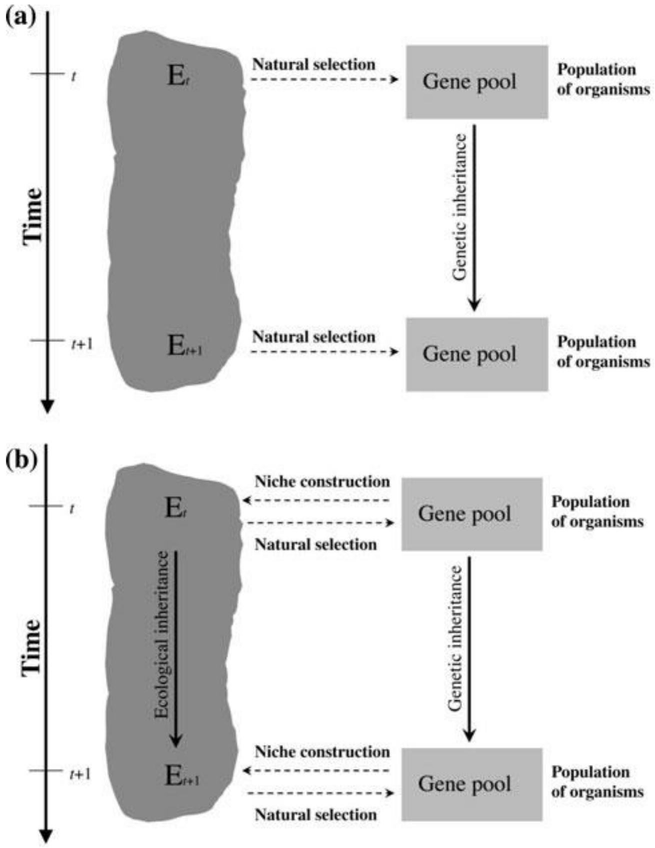


Figure 3.12. (a) Standard evolution theory. (b) Niche construction theory. Reproduced from Laland and O'Brien (2011, p. 192). There are two views of evolution. Under the conventional perspective (a), niche construction is recognised as a product of natural selection.

According to Laland et al. (2016), the term ecological inheritance was first used by Odling-Smee to establish NCT and its connection with cultural inheritance, which is the knowledge and material artefacts acquired by an organism during its lifetime. In niche construction, an organism changes its environment and the changed environment feeds back to the organism.

Wilson and Clark (2009) opine that although niche construction has long been recognised, the works of Odling-Smee et al. place niche construction in centre stage. NCT introduces a missing dynamic aspect to SET, i.e. the importance of ecological inheritance resulting from the modification of the environment by the organism. NCT integrates ecology and genetics via an extended notion of ecological inheritance, i.e. inheritance via genetic and non-genetic processes.

Odling-Smee et al. (2003) claim that NCT accounts for ecological inheritance, i.e. inheritance via an external environment in addition to the genetic inheritance of evolutionary theory. The combined inheritance is called niche inheritance, and it is used to establish the relation between evolution and development.

Although NCT puts the dynamic interactions and modifications of the environment by an organism at centre stage, there are criticisms of such a conceptualisation of NCT. One of the notable criticisms is from Dawkins (2004), who argues that this conceptualisation of NCT is misleading. He claims that the extended phenotype can capture how the niche changes. A phenotype is an organism’s observable traits. Dawkins argues that an allele (a DNA variant due to mutation) can produce an extended phenotype, which can capture the environmental modifications made by an organism and their effects on niches. Dawkins argues that rather than using niche construction, we must focus on changes to a niche. Moreover, Gupta et al. (2017) argue that NCT already encompasses SET and that there was no requirement for a separate theoretical formulation of NCT. Whatever the case, the proponents and opponents of NCT agree that the environmental modifications by an organism affect evolutionary pressures. The disagreement is about how to capture the changes made to the environment by the organism and ecological inheritance. The proponents of SET prefer to capture these aspects in a gene-centric model. Dawkins prefers the extended phenotype in SET. However, the proponents of NCT conceptualise the interactions between the organism and its environment, the changes to the environment by the organism, and the resulting ecological inheritance to explain evolutionary process.

Odling-Smee and Turner claim that NCT recognises the changes made by organisms to their environment and their impact on the ecological inheritance of evolutionary processes (Odling-Smee & Turner, 2011). Laland and Sterelny (2006) argue that most evolutionary biologists neglect the causal role of an organism in evolution and the multigenerational feedback from an extended phenotype. Wilson and Clark (2009, p. 8) argue that being key to survival, “the relevant *phenotype* should not always be sought solely within the organismic skin bag”.

NCT considers how both individual organisms and species shape their environments. NCT, indeed, acknowledges the significant role that individual organisms have in shaping their environments through their activities and choices. This shaping, in turn, affects the evolutionary pressures faced by these organisms and their offspring. The individual organisms, through their unique behaviours and interactions with their environment, contribute to ecological changes. This individual-level focus is crucial for understanding the nuances of NCT, as each organism’s actions can have a cumulative effect on the species and the ecosystem. A feedback loop in NCT is more nuanced and specific compared to a feedback loop in SET, as it describes how individual organisms, through their activities and choices, can change their environment in ways that affect their own evolutionary trajectory. This perspective acknowledges that an organism’s behaviour and life-history strategies are as much a part of its evolutionary narrative as are its physical characteristics. Although the extended phenotype (originally proposed by Richard Dawkins) does consider the impact of an organism’s actions on its environment, NCT provides a broader framework. It not only encompasses the immediate effects of an organism’s phenotype on its surroundings but also how these changes feed back into the evolutionary process over time.

From my perspective, what is important here is that a feedback loop between an organism and its environment changes the environment. This, in turn, results in ecological inheritance, which acts as an evolutionary process in its own right.

The following section explores the processes of human niche construction.

3.5.3 Human Niche Construction

Laland and O’Brien (2011) explain niche construction as a process in which “organisms, through their activities and choices, modify their own and each other’s niches” (Laland & O’Brien, 2011, p. 191), just like a dam made by a beaver can affect microorganisms, plants and other animals. Constructing a niche transforms natural selection pressures and feeds back to evolution at various levels. Niche-constructing species can engineer their ecosystem by manipulating their environment, and the acquired characteristics can evolve through natural selection. This is particularly relevant for human evolution, since humans extensively modify the environment through their cultural practices.

Laland and O'Brien (2011) state that humans can acquire knowledge through a set of information-acquisition processes operating in three different domains:

(1) Population genetic: Genetic information can be transferred through inheritance.

(2) Ontogenetic: Information can be acquired through ontogenetic processes such as learning.

(3) Cultural:

> It is readily apparent that contemporary humans are born into a massively constructed world, with an ecological inheritance that includes houses, hospitals, farms, factories, computers, satellites, and the World Wide Web. Niche construction and ecological inheritance are thus likely to be at least as consequential for developmental processes as they are now known to be in human evolution. (Laland & O'Brien, 2011, p. 195)

Figure 3.13 shows the set of processes that humans can use to acquire information. Three domains feed back into niche construction: population genetics, ontogeny and culture. The domains are distinct but interconnected.

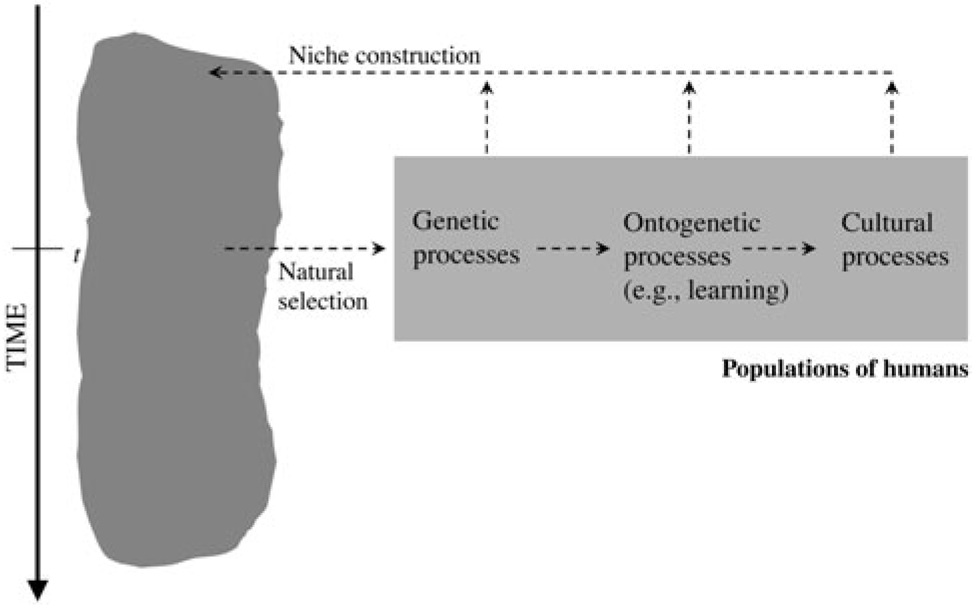


Figure 3.13. Processes used to acquire information. Reproduced from Laland and O'Brien (2011, p. 195).

Laland and O'Brien (2011) argue that:

> The three domains are distinct but interconnected with each interacting with, but not completely determined by, the others [Figure 3.14]. That is, learning is informed by, but not fully specified by, genetic information, and cultural transmission may be informed by, but again, not completely specified by, both genetic and developmental processes. Genes may affect information gain at the ontogenetic level, which in turn influences information acquisition in the cultural domain. In addition, ontogenetic processes – particularly learning – may be affected by cultural processes, whereas population-genetic processes may be affected by both ontogenetic processes and cultural processes when humans modify environments, generating selective feedback to each process.
>
> (Laland & O'Brien, 2011, p. 195)

Figure 3.14 shows the interconnections.

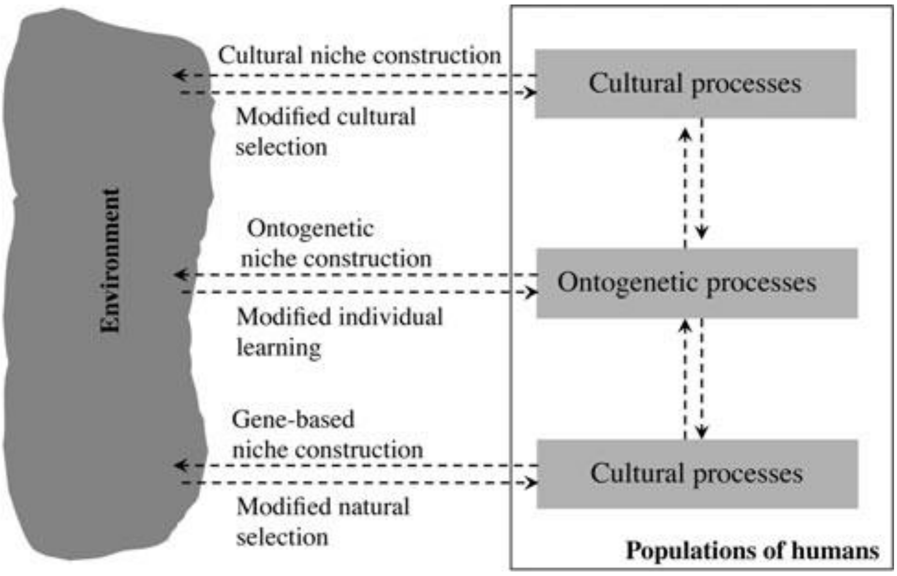

Environment
Cultural niche construction
Modified cultural selection
Ontogenetic niche construction
Modified individual learning
Gene-based niche construction
Modified natural selection
Cultural processes
Ontogenetic processes
Cultural processes
Populations of humans

Figure 3.14. Interconnections between information processes and the environment. Reproduced from Laland and O'Brien (2011, p. 196).

Human niche construction, through modification of the environment, creates artefacts and other ecologically inherited resources that not only act as sources of biological selection on human genes (Laland et al., 2011) but also facilitate learning and mediate cultural traditions. Much of human niche construction is guided by socially learned knowledge and cultural inheritance; however, the transmission and acquisition of this knowledge is itself dependent on pre-existing information acquired through genetic evolution, complex ontogenetic processes or prior social learning (Laland & O'Brien, 2011, p. 197).

An example of human niche construction is evident in the recent COVID-19 pandemic. Widespread urbanisation (such as the construction of villages, towns and cities) and advanced transportation (such as planes, ships and automobiles) have made the world a global village. Such huge and dense population centres have created new hazards, such as the rapid spread of diseases. People are now interconnected intercontinentally. Thus, just after it arose, COVID-19 became a pandemic without any geographical boundaries. Based on the three domains in Figure 3.14, genetic, ontogenetic and cultural, how did humanity respond to this situation? Humans responded to the novel selection pressures through a cultural revolution. Cultural selection occurred by manufacturing vaccines, masks, hand sanitiser and ventilators and by constructing hospitals. At the ontogenetic level, infected people developed antibodies, which may confer some immunity. At the genetic level, through biological evolution, a resistant genotype may be selected for. It is clear that, as Odling-Smee et al. (2003) argue, cultural niche construction with culturally acquired information offers a more immediate solution to new challenges.

According to Odling-Smee and Laland (2011), NCT includes the non-genetic inheritance of culturally acquired information or knowledge and artefacts, which can modify the selective environments of organisms within and between generations. "Human cultural niche construction is also a cause of changed developmental environments and sources of modified selection to a multitude of other species that inhabit the human niche" (Laland & O'Brien, 2011, p. 199).

3.5.4 EC and NCT

Regarding EC, Clark and Wheeler argue that: "Under certain conditions, non-organic props and aids, many of which are either culturally inherited tools or structures manipulated by culturally transmitted practices, might themselves count as proper parts of extended cognitive processes" (Wheeler & Clark, 2008, p. 3564). Here EC coincides with NCT to yield cognitive niche construction. NCT describes how organisms actively shape their environments, which can influence subsequent generations and their evolutionary trajectories. This idea is aligned with EC, as the modifications that organisms make to their environment can be seen as extensions of their cognitive processes. The use of external tools or structures plays a crucial role in both NCT and EC. In NCT, organisms engaging with the environment create physical structures and modify landscapes, which is a form of cognitive extension. Niche construction activities can be seen as a form of EC, as organisms use their environmental modifications as cognitive tools to enhance their adaptive capacities. Conversely, the changes that organisms make to their environment through niche construction can shape the cognitive demands and opportunities for subsequent generations, potentially leading to further evolutionary shifts. As we create artefacts that extend our cognition, we often then further adapt our environment in response to our use of those artefacts.

3.5.5 Social Niche Construction

Social niche construction can explain sociality. Constructed cultural niches include artefacts that are passed on to future generations. These can significantly alter the subsequent course of human evolution:

> In NCT, "culture" includes the non-genetic inheritance of culturally acquired information, or "knowledge" and of material culture, or "artefacts" both of which can modify the selective environments of the organism, within and between generations. For example, the introduction of artefacts in 20th

> century is most unlikely to depend upon human genetic changes. However, the cultural innovation clearly feeds back to influence subsequent human cultural activities. (Odling-Smee & Turner, 2011, p. 286)

Laland et al. (2016) argue that: “Social niche construction describes the situation in which individuals, singly or collectively, influence the composition and dynamics of their environment” (Laland et al, 2016, p. 199).

Odling-Smee and Turner (2011) claim that a strictly genetic approach in the context of NCT will not be sufficient to explain social niche construction since it requires a cultural evolutionary component.

Figure 3.15(a) shows the relation between NCT and culture. From the expression of the genotype (G), culture (C) is derived, which can develop and modify the ecological inheritance for future generations. Artefacts are a simple form of cultural inheritance for future generations. Figure 3.15(b) shows cultural inheritance as a component of ecological inheritance, which comprises the inheritance of material artefacts and socially transmitted cultural knowledge.

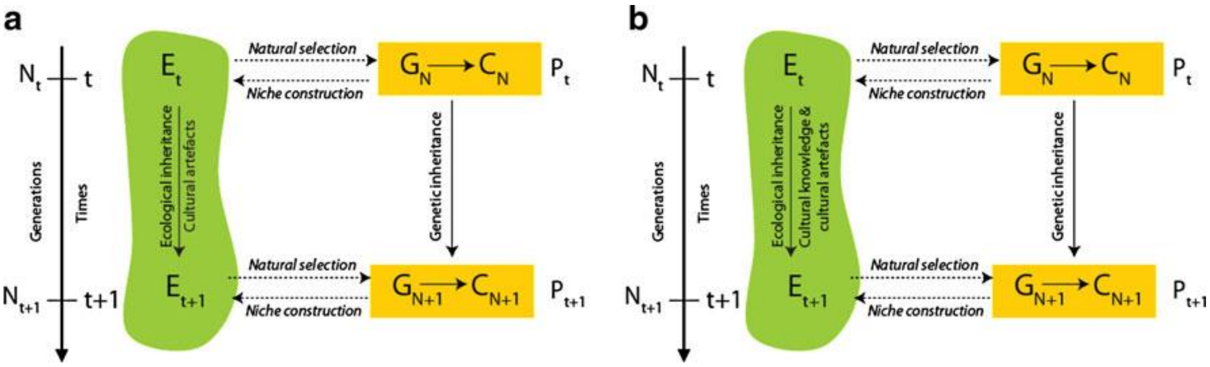


Figure 3.15. (a) NCT and culture. (b) Cultural inheritance. Reproduced from Odling-Smee and Turner (2011, p. 287).

The difference between the figures is that in Figure 3.15(b), cultural artefacts are included along with cultural inheritance. An example of Figure 3.15(a) is the architecture created by colonies of termites and other social insects.

As we can see, SET cannot account for or explain Figure 3.15(a), as genetic inheritance does not provide a comprehensive description of the evolution of social systems. What it can explain, however, is the complex flow of information, energy and matter that ensures the collective behaviours have sufficient coherence to enhance the persistence and ultimate reproduction of the collective of a superorganism.

Figure 3.15(b) shows that culture in one generation endures sufficiently long to influence the selective environments of future generations, not through genetic inheritance and not solely through the niche constructed by a cultural artefact, but also through directly transmitted knowledge (Odling-Smee & Turner, 2011, p. 286).

As we have seen, NCT emphasises the dynamic interplay between an organism and its environment. It underscores how organisms, through their activities and behaviours, actively shape their ecological niches. This shaping is not just physical but also involves the creation of social and cultural dimensions within these niches.

3.5.6 Cognitive Niche versus Cognitive Niche Construction

In the previous sections, we discussed NCT, which explains genetic and ecological inheritance as evolutionary selection pressures, compared to the solely genetic inheritance of SET. It is hard for evolutionary biologists to explain the surprising cognitive development of *Homo sapiens* compared with our next of kin in the evolutionary ladder. Humans adapt better to novel environments than other organisms. In the past few decades, cognitive niche theories have been developed to answer this question. Bertolotti et al. (2017) state that cognitive niche theories are bridging evolutionary biology, philosophy, cognitive science and anthropology through an interdisciplinary approach. However, they notice that there are multiple types of cognitive niche theories. Two competing theories attempt to explain the cognitive development of humans from an evolutionary perspective:

1. Cognitive niche theory (CN)

2. Cognitive niche construction theory (CNC)

CN provides an evolutionary explanation for the cognitive development of humans by considering improvisational intelligence using cause-and-effect reasoning (Tooby & DeVore, 1987; Pinker 2010, 2014). In CN, cognitive evolution is considered to be a form of co-evolution along with genetic evolution. In this co-evolution, selection for enhanced cognition is engaged in a positive feedback loop. A positive feedback loop in evolution refers to a process in which an evolutionary change leads to consequences that further accelerate or enhance that change. In the context of cognitive evolution, as humans developed better cognitive abilities, these abilities improved their survival and reproductive success, which in turn further increased the selection pressure for even better cognitive abilities. In CN, cognitive evolution is seen as a co-evolutionary process with genetic evolution, such that improvements in cognitive abilities offer survival and reproductive advantages, which in turn drive further cognitive development. A cognitive niche is an environment in which the survival and reproductive success of an organism are heavily dependent on its cognitive abilities, particularly its understanding and manipulation of the environment through cause-and-effect reasoning. In such a niche, the ability to reason, plan and solve problems is a crucial evolutionary trait. This type of environment promotes the development and refinement of these abilities. Thus, cognitive evolution within this niche is inherently linked to the development of complex cognitive skills, such as improvisational intelligence. The development of improvisational intelligence and sophisticated cause-and-effect reasoning feeds into this loop. As humans became better at solving complex problems, they could manipulate their environment more effectively, leading to new challenges and opportunities that further selected for enhanced cognitive abilities. A cognitive niche provides the context and the selective pressures that drive the evolution of these skills. Human intelligence results from selection for improvisational intelligence, i.e. intelligence capable of generating complex solutions to novel problems. Cognitive evolution occurs in a cognitive niche using cause-and-effect reasoning.

Boyd et al. (2011) claim that, according to CN, humans have evolved improvisational intelligence with flexible cognitive capacities that allow humans to acquire locally adaptive behaviour in a wide range of environments. Humans are adapted to their cognitive niche and these capacities are augmented by our ability to learn from each other, especially using grammatical language. However, animals are limited to domain-specific learning and decision-making that is adapted to a particular environment (Boyd et al., 2011, p. 10198).

Following NCT, Clark (2005) and Sterelny (2003, 2006) introduce CNC to explain the cognitive development of humans. As the name indicates, cognitive niche construction theory is based on NCT, which involves niche construction by an organism in its environment and the evolutionary pressure from the changed environment on the organism. However, in *cognitive niche* theories, the environment of the cognitive agent is often taken to be unresponsive to the actions of the agent. These two theories – CNC and CN – have many commonalities, although the basis of the theories differs.

CNC provides an insight into the evolutionary pressures from ecological inheritance. Clark defines cognitive niche construction as the process by which animals build physical structures that transform problem spaces in ways that aid (or sometimes impede) thinking and reasoning about some target domain or domains. These physical structures combine with appropriate culturally transmitted practices to enhance problem-solving and, in the most dramatic cases, to make possible whole new forms of thought and reasoning (Wheeler & Clark, 2008, p. 3564). Wilson and Clark (2009) argue that thinking is a kind of niche construction: "Thinking is a kind of building, a kind of intellectual niche construction that appropriates and integrates material resources around one into pre-existing cognitive structures" (Wilson & Clark, 2009, p. 61).

Sterelny (2012) notes that human niche construction activities accumulate, especially with the creation of learning environments: "In the case of humans much niche construction is the 'epistemic engineering' of the informational character of agents' environments" (Sterelny, 2003, p. 147). The cultural transmission of information gradually becomes cumulative:

> Humans are downstream niche constructors par excellence. One important aspect of that niche construction is altering the epistemic environment of our offspring. We engineer the informational environment of our downstream generation, thus making for more accurate and reliable acquisition of key capacities. Teaching is a form of downstream niche construction. It is one way developmental environments can be engineered to overcome the frailties of one to one informational transmission. (Sterelny, 2006. p. 154)

There is some interconnection between CNC and CN. CNC has two relational arrows, i.e. the arrow from the organism to the environment and the arrow from the environment to the organism. However, CN considers only the arrow from the organism to the environment. The "construction" part of cognitive niche construction, i.e. the effect of a change to the environment by the organism and the feedback loop with the organism, is missing in CN. Except for the construction part, there are commonalities between CNC and CN. Kerr (2007) argues that:

> It is simpler to deal with a *one-way* causal arrow flowing from environment to agent. … In this view, it is the autonomous properties of the environment that explain the cognitive properties of the agent. Niche construction complicates this simple causal picture by making the properties of the environment partially dependent on properties of the agent. A causal loop is

> introduced, where an organism is both affected by *and affects* the environment. (Kerr, 2007, p. 259)

There are a few gaps in correlating NCT to cognitive development, such as how ecological inheritance causes intelligent behaviour. Since the focus of this section is to provide an independent argument to support the feedback loop between the cognitive agent and the environment, I will unpack the nature of engagement of humans with the environment.

Bertolotti et al. (2017) use enablement (a niche enables the survival of an organism) and affordance (adaptive fit, i.e. the suitability of the environment) to connect NCT to cognitive niche construction. Bertolotti et al. (2017) argue that ecological niche construction (NCT) shifts into cognitive niche construction because enablements shift into affordances. They argue that using the biological notion of enablement and the psycho-cognitive notion of affordance, the theoretical overlaps of ecological (NCT) and CN can be addressed:

> The most coherent way to express something close to causality in biological systems, according to Longo and Montévil, is through the notion of *enablement*, which accounts for the changes of "phase space" broadly meant, namely for the emergence of a new, previously uncomputable observable. In short, a niche *enables* the survival of an otherwise incompatible/impossible form of life, it *does not cause* it. … The notion of enablement is not presented as a system-dependent feature (such as physical causation, whose laws affect the whole system they refer to), nor as a species or individual-dependent trait, but as an interaction-dependent feature. This is akin to how niches, in the constructionist view, are constructed by the organisms by negotiation with the environment, in the form of a continuous feedback circle.
>
> (Bertolotti et al., 2017, p. 4770)

Affordance and adaptive fit are two of the core aspects of human niche construction. To explain perception and the role of the optic array, Gibson (1979) introduces invariants and affordance. Cognitive processes based on affordance enable us to engage with society and the world. Gibson's concept of affordance gives an insight into the cognitive activities of an agent and their interaction with the environment that enables the agent to survive. Gibson argues that a cogniser is not passive (as considered by the computational theory of cognition). Instead, a cogniser is active, and the motion of their body and its interaction with the environment are constituent parts of the cognitive process. Affordance relates to the potential utility of an object in the environment, e.g. it indicates whether a chair is suitable to sit on. Affordance has agent relativity, such that the perception of the affordance of a chair by a lean man may suggest it is suitable to sit on, but for a larger man, affordance may indicate that the chair is not suitable to sit on.

Bertolotti et al. (2017) argue that "the respective cores we individuated, enablement and affordance, may act as a conceptual fulcrum making sense of the similarities between the two theories: ecological niche construction shifts into cognitive niche construction because enablements shift into affordances" (Bertolotti et al., 2017, p. 4779).

I conclude that NCT highlights the importance of the feedback loops in cultural and ontogenetic niche construction, which supports cognitive niche construction and the feedback loops between external artefacts and the agent in cognition.

In the next section, I focus on patterns and pattern recognisers in general (especially on patterns in external information-bearing structures) and on the histories embedded in external information-bearing structures to provide an independent argument for the constitutive role of an external artefact due to the feedback loop and the manipulation of external information-bearing structures by a cognitive agent.

3.6 Argument from the Patterns in External Information-bearing Structures

What is a pattern? A pattern can be abstract, like concepts such as the centre of gravity or the periodic movement of a pendulum, or physical, like an array of tiles. As we have seen, the application of dynamic principles highlights patterns in behaviour, such as periodicity or chaotic behaviour. In Section 3.3, I described the application of DST to cognition to identify cognitive and behavioural patterns. In Section 3.3.2, I explored how DST is applied to understand cognitive and behavioural patterns. For example, DST helps explain simple motor behaviours, like finger wagging as discussed by Kelso (1995), and cognitive errors in children, such as the A-not-B error described by Thelen and Smith (1994). These theories view cognitive and behavioural patterns as part of a larger system that follows certain rules or "laws". According to DST, these patterns tend to be predictable and often return to a stable state after a disturbance. This means that behaviours and cognitive processes can be seen as deterministic to some extent, as they follow predictable paths when influenced by certain conditions or stimuli.

Dennett (1991) suggests that for a pattern to be real, it should be compressible. This means that the pattern can be described or summarised in a way that requires less information than describing every single detail (a bit-by-bit duplication). He gave a second criterion of predictability. A pattern is considered real if it allows predictions to be made. If we can use the pattern to anticipate future outcomes or behaviours based on current or past data, then the pattern has real-world applicability and significance. For instance, if observing a certain weather pattern allows meteorologists to predict rain accurately, then this pattern is real according to Dennett's criteria. Thus, a pattern is real if it can be used to make accurate forecasts or predictions about future events or states.

I consider three aspects of the EM thesis: (1) the cogniser, (2) the external artefact or information-bearing structure and (3) the dynamic interaction between the cogniser and the external artefact.

By evaluating various examples supporting the EM thesis, such as Otto (notebook), it is clear that EM theorists have overlooked the nature and characteristics of external information-bearing structures. In line with Dennett's claim about real patterns in the world, such real patterns are associated with information-bearing structures and each pattern has a history. For example, pencil and paper and the internet have their own histories and patterns. Dennett (1991) has a mild realism about the ontology of patterns. Such patterns are real. They are abstract objects just like the centre of gravity and electrons in science, which are close to reality. Therefore, they can be considered to be real. Applying Dennett's criteria to the EM thesis, we can argue that the patterns associated with external cognitive aids, like a notebook, are indeed "real". They are not just abstract concepts; they meaningfully represent and summarise complex information and help in predicting or understanding cognitive processes.

A pattern requires a pattern recogniser. As Dennett states: "In the root case a pattern is 'by definition' a candidate for pattern recognition" (Dennett, 1991, p 32). The patterns embedded in external information-bearing structures are independent of the pattern recogniser. Dennett's assertion that a pattern essentially requires a recogniser reflects the idea that patterns are not inherently "meaningful" or "recognisable" without an entity (like a human or a machine) to interpret them. Such an interpretation involves identifying and understanding the pattern, thereby giving it significance or meaning. If the patterns embedded in external information-bearing structures are independent of the pattern recogniser, then the focus shifts to how these patterns are formed and evolve. Such patterns emerge through the cumulative processes of human knowledge acquisition and innovation. They exist as structured data or information sequences, regardless of whether they are actively being recognised or interpreted at any given moment. Patterns in external structures, like written language, mathematical formulas and digital data, do exist independently because of human activity and evolution. However, for these patterns to be meaningful or to serve a purpose, like conveying information or solving a problem, they require a recogniser who can interpret and understand them. A pattern recogniser can recognise a pattern based on their skills and acquired knowledge. However, how this happens will be different when, for the first time, the agent identifies a new pattern, as when somebody invents something or develops a scientific theory.

3.6.1 Patterns in Information-bearing Structures

Are there patterns in external information-bearing structures? If there are, what are their characteristics? I argue that there are patterns embedded in external information-bearing structures and that they were developed by humans via niche construction. As detailed in Section 3.5.6, CN and CNC support the patterns embedded in external information-bearing structures, such as artefacts. From the historical perspective, it is evident that humans have explored and invented various artefacts with increasing complexity to offload cognition, such as a pen and paper, the printing press, computers, the internet, microscopes, telescopes etc. As an organism develops or grows, new patterns emerge via self-organisation. As humans interact with their environment and create new tools, they inadvertently contribute to a process in which the patterns in these tools emerge and reorganise over time, often without a central guiding hand. This self-organisation is a key characteristic of how artefacts develop and become more complex, and it reflects the increasing sophistication of human cognitive abilities and societal needs. DST provides a framework for understanding and analysing these evolving patterns. By applying DST, we can examine how various elements within these systems interact over time such that new patterns emerge or existing ones are modified.

In the context of NCT, CNC and CN, human-made artefacts evolve over generations to become increasingly complex. This evolution is driven by the accumulation of knowledge and skills, which is facilitated by the human ability to construct and modify cognitive niches. This process is further enhanced by humans' improvisational intelligence and their understanding of cause-and-effect relationships, which enable them to innovate and refine their tools and technologies effectively. From an evolutionary perspective, as outlined in theories like NCT, CNC and CN, artefacts (tools, technologies, etc.) evolve over time. This evolution refers to how these artefacts increase in complexity across generations. Thus, “evolution” does not just mean biological evolution but also includes the development and refinement of the tools and technologies that humans use. Over time, these artefacts become more complex and sophisticated as they are continually adapted to the changing needs and capabilities of human societies. The term “cumulative epistemic engineering” refers to the gradual accumulation of knowledge and skills over generations, which leads to advances in technology and understanding. This process is a key component of cognitive niche construction, as humans not only adapt to their environment but also actively modify it using their knowledge and tools. In this context, cognitive niche construction is the process by which humans create and shape their environment – including their artefacts – through their cognitive abilities, such as problem-solving, planning and innovation. Improvisational intelligence is the ability to devise creative and effective solutions to new and complex problems. It is a key aspect of human cognition, and it enables us to innovate and improve our tools and technologies. Cause-and-effect reasoning is a cognitive process that allows humans to understand the relationship between their actions and the consequent outcomes. This type of reasoning is crucial for developing and refining artefacts. By understanding how different modifications affect the functionality of a tool, humans can make informed improvements, which drives the evolution of more complex artefacts.

From the NCT perspective of artefacts, “humans are massive constructors of developmental environments. By modifying the world, human niche construction creates artefacts and other externally inherited resources that not only act as sources of biological selection on human genes but shape the learning opportunities and developmental trajectories of recipient organisms” (Flynn et al., 2013, p. 299). Wheeler and Clark (2008, p. 3564) asset that “cognitive niche construction” occurs when “animals build physical structures that transform problem spaces in ways that aid (or sometimes impede) thinking and reasoning about some target domain or domains”. The physical and informational legacies “make possible whole new forms of thought and reason”.

A pattern recogniser’s engagement with external information-bearing structures goes through various phases. The initial engagement may have feedback loops, in which the external information-bearing structures are manipulated to accomplish cognitive tasks. Depending on the nature of the engagement, various loops are possible with an external artefact. We must consider two aspects regarding the nature of the engagement between a pattern recogniser and an external artefact:

1. The history of the external information-bearing structure: Pencil and paper have a long history of being used to carry out epistemic actions. Pencil and paper, as external information-bearing structures, have a significant history in facilitating epistemic actions. Although their basic use for tasks like simple note-taking or basic arithmetic may not demand extensive skills, more complex applications, such as making detailed drawings or advanced mathematical calculations, indeed require a considerable amount of acquired skill and practice. These tools have played a vital role in enabling individuals to offload cognitive tasks, effectively extending their cognitive capabilities. Pencil and paper enable the agent to offload part of her cognitive load. In human history, there are many examples of external artefacts being used to offload some cognitive activities. Before the development of a full-fledged combinatorial language, early humans used pictures to offload some of their cognitive activities. Early agrarian societies used signs and pictures to quantity and remember agricultural produce and the cost of the labour involved. However, it is important to note that the skills required to interact with and derive benefit from these tools vary greatly. Whereas pencil and paper may require relatively basic skills for simple tasks, other tools, like software or night vision goggles, demand highly specialised skills to enable pattern recognition. The evolution of these artefacts is marked by increasing complexity, both in their design and in the cognitive skills required to utilise them effectively. This progression reflects the dynamic relationship between human cognitive development and the tools we create and use. The patterns associated with artefacts have evolved and become increasingly complex.
2. The nature of the interaction between the agent and the external artefact: The nature of this engagement can differ. Some information processing, i.e. the identification of a pattern by a cogniser, may be an open loop, i.e. there is no requirement for the agent to manipulate external information-bearing structures when exploring patterns. I argue that offloading cognitive activities to an external artefact is one form of engagement. Pencil and paper are an example of this. However, some cognitive tasks that are accomplished by manipulating an external artefact require continuous reciprocal causation (CRC) between the agent and the external artefact. The use of night vision goggles by a soldier is one such example.

3.6.2 History of External Information-bearing Structures

In the preceding section, we examined the patterns embedded within external information-bearing structures and the role of these patterns in cognitive processes. This section describes the historical evolution of such artefacts. It evaluates the historical trajectory of external information-bearing structures by examining three key areas: (1) developments in conceptual abstraction and invention, (2) cultural inheritance and cognitive patterns and (3) the interactions between social institutions and conceptual abstraction.

From the historical evidence, Basalla (1988) argues that there is a continuity in the evolution of artefacts. Contrary to the traditional conception of necessity as the mother of invention, the emergence of new artefacts has to be evaluated based on continuity, novelty, diversity and selection, as explained earlier. The history of patterns embedded in external information-bearing structures from any period in time should be evaluated based on:

1. Developments in abstract concepts, such as new inventions, and the conceptual development of a period: The abstraction of concepts occurs when we identify a new pattern, such as the centre of gravity. Along with the continuity in the emergence of new artefacts, the abstraction of concepts, which can be the result of improvisational intelligence based on cause-and-effect reasoning, occurs as outlined in CN. According to CNC, the continuity in the emergence of novel artefacts can be based on cultural inheritance and subsequent cognitive niche construction. How is a new artefact created? Often, when science connects with technology.

2. Cultural inheritance, cognitive practices and cognitive patterns across society during a period: Throughout history, we can identify progress in learning, the assimilation of inventions and the emergence of novel abstract concepts or the identification of new patterns. When a new invention or an abstract concept is learned by the majority of people, it becomes a cognitive pattern across society. Cognitive practices are patterns across society. The cognitive processes used by an individual can extend into the environment, and they can become transferred to another individual and become part of that person’s own cognitive processes without requiring a group or collective mind. As Sterelny (2001) argues, social learning is unique to humans. Language also plays an important role in social learning.

Combinatorial grammatical language has a key role in social niche construction and cooperation among people. It allows an individual to interact with society and with their environment through social affordance, technical know-how and tool use. Cognitive niche construction, social niche construction and their interaction with the environment show how social learning enables cognitive practices, since the majority of people can learn and assimilate abstract concepts. The capacity of people to learn and understand new concepts already identified is called psycholinguistic metaphorical abstraction (Pinker, 2010).

Similarly, Menary (2010) argues that cognitive capacities are extended by socio-cultural practices. Cognitive practices, such as mathematical cognition, are patterns of activity that can spread throughout a population. For example, it was commonly accepted that the Earth was flat until this belief was replaced by the notion of a spherical Earth. When Einstein originally put forward the concept of relativity, few people could understand this abstract concept of space–time, though as Stephen Hawking points out, millions can now understand the theory of relativity. This shows how abstract concepts can be learned through metaphorical abstraction.

3. The dynamic interaction of social institutions and the continued abstraction of concepts: The wider population learn abstract concepts through metaphorical abstraction. Social institutions then accommodate the new cognitive patterns and cultural practices. Changes to social institutions occur when newly learned practices cannot be accommodated within the existing structures. For example, in physics, the Newtonian deterministic world outlook was replaced by a probabilistic world outlook with the development of relativity and quantum mechanics.

3.6.3 Conclusion

In the discussions presented in Sections 3.4, 3.5 and 3.6, I have argued that dynamic interactions, such as feedback loops and the manipulation of external information-bearing structures, play a crucial role in how external artefacts contribute to cognition. Although I theorised that these interactions are both necessary and sufficient for the constitutive role of external artefacts in cognition, it is important to consider the broader context provided by recent developments in various fields.

The application of dynamical systems theory (DST) to cognition, along with insights from evolutionary biology through niche construction theory (NCT), cognitive niche theory

(CN), cognitive niche construction theory (CNC) and developmental systems theory, generally aligns with the principles of the extended mind (EM) and extended cognition (EM) hypotheses. These fields emphasise the importance of both internal and external factors in shaping cognitive processes. However, their focus is often on broader organism–environment interactions over evolutionary or developmental timescales, rather than on specific artefacts. Therefore, although these theories support the general idea that external factors influence cognition, they may not specifically validate the claim that feedback loops and the manipulation of external artefacts are alone necessary and sufficient for cognitive extension.

The explanatory power to account for the historical progress of abstract thought and the impact of abstract thought on the development of tools, artefacts and social institutions and on cognition should be considered as one of the criteria needed to establish the role of external features in cognition and mental states.

From the previous discussions, it is clear that NCT not only substantiates the importance of feedback loops with the environment but also provides an explanation for the cognitive development of humans from an evolutionary perspective, such as in cognitive niche construction. The application of dynamics to cognition not only supports the necessity for feedback loops and the constitutive role of external artefact but also illustrates the nature of cognitive agents over time. Developmental systems theory takes a systems approach to cognitive development and the nature of a cognitive agent. Patterns in external information-bearing structures not only explain the nature and histories associated with the engagement of an external artefact by a cognitive agent, but they also explain the continuity in the development of artefacts and the continuity and progress in the abstract reasoning embedded in the development of external artefacts. Such explanations provide historical insights into the cognitive development of humans as well as the nature of the relations between a cognitive agent and artefacts.

3.7 Modified EM

As can be seen from the various theories, such as NCT, DST and developmental systems theory, and from patterns in external information-bearing structures, feedback loops are versatile. Our main focus is the participation of external information-bearing structures in feedback loops. Thus, our modified version of EM is based on feedback loops involving an agent and an external information-bearing structure that has a pattern or history associated with it.

I argued in Section 3.3.4 that dynamic interactions, such as feedback loops with external information-bearing structures, and the manipulation of external information-bearing structures are sufficient and necessary conditions for the constitutive role of external artefacts in cognition. If the cogniser and external artefact both have an active role in cognition via a feedback loop, I argue that the overall system comprising the cogniser, the external information-bearing structure and the feedback loop are the constituent parts of cognition. This role of an external information-bearing structure via a feedback loop cannot be considered to be an enabling role. An enabling role implies that the external structure merely facilitates or aids cognitive processes without being an integral part of those processes. In contrast, a constitutive role suggests that the external structure is an essential and inseparable component of the cognitive process. The distinction hinges on whether the cognitive process can be fully realised without the involvement of the external structure. If the process depends fundamentally on the structure, then the role of the structure is constitutive. Feedback loops involving external information-bearing structures and agents (cognisers) are dynamic and interactive. In such loops, the structure and the agent are continuously influencing and modifying each other's states. This mutual influence implies that the external structure is not just aiding the cognition but actively shaping and being shaped by the cognitive process. It is a participant in the cognitive activity not just a facilitator. If we consider only the enabling role of an external artefact in cognition and undermine the feedback loop with the cogniser, then we lose the dynamic interaction of the cogniser in the manipulation of the external artefact and the reciprocal causation of the manipulated external information-bearing structure on the cogniser. Further, we lose the changes in the cognitive processes that occur over time due to the interactions between the external information-bearing structure and the cogniser via the feedback loop. Furthermore, in such cases, the cognitive task cannot be accomplished without the feedback loop. Therefore, in situations with a feedback loop, especially when an agent manipulates an external information-bearing structure, I cannot see any reason why we cannot consider the external artefact as a constituent part of cognition. The enabling role of an external artefact is captured well in the example of pen and paper. The agent offloads some of their cognitive load to these external artefacts. In this case, there is no feedback loop and the external artefact is not manipulated to achieve cognitive success. The enabling role of an external artefact can be considered as an open loop, such as in the example of pen and paper. However, the manipulation of night vision goggles cannot be explained by the enabling role of the goggles, as this neglects how the agent manipulates the goggles to achieve cognitive success. As Hurley (2010) states, the explanatory potential has to be considered case by case when establishing whether an external artefact has an enabling or a constitutive role in cognition. The constitutive role of an external artefact via a feedback loop and the manipulation of that artefact by the agent to achieve cognitive success is better aligned with NCT, CNC, CN, developmental systems theory and the application of DST to cognition.

In summary, EM requires the following:

- An external information-bearing structure that has associated patterns or a history. This requirement suggests that if an external structure is to be considered part of a cognitive process in the EM framework, it cannot be just any physical object or tool. It needs to have associated patterns or a history, which means that it contains or represents information in a structured way that has evolved or been developed over time. The associated patterns refer to the structure's ability to hold or process information in a recognisable and usable form. This could be as simple as lines written on paper or as complex as the software in a digital device. These patterns are crucial for the structure's utility in cognitive processes. A history implies that the structure has been shaped or refined through human use and cultural evolution. This historical aspect is significant because it means that the structure has been integrated into human cognitive practices over time, making it more likely to be effectively utilised in cognitive processes.
- An agent who establishes a feedback loop that results in the manipulation of the external information-bearing structure to accomplish a particular epistemic task.

Since external information-bearing structures play a constitutive role in the cognitive processes of the agent, then cognition is extended. Consequently, the beliefs formed by that cognition are extended, i.e. the mind is extended.

3.7.1 Critical Evaluation of the Modified Version of EM

If a cognitive process has a feedback loop with an external artefact that results in the manipulation of that external information-bearing structure, then clearly this external artefact has a constitutive role rather than an enabling role. In a feedback loop, the agent acts on the external artefact and manipulates it to accomplish cognitive success, such as in the manipulation of night vision goggles to obtain an image of an object at night. Beliefs are formed by this two-way interaction between the agent and the external artefact. I argue that feedback loops involving the manipulation of external information-bearing structures are sufficient and necessary for EM. The theory is immune from the coupling-constitution fallacy and cognitive bloat. The coupling-constitution fallacy occurs when there is confusion between an external artefact that is merely coupled (connected or associated) with a cognitive process and one that is a constitutive part of that process. In standard EM discussions, critics argue that just because a cognitive process is coupled with an external artefact, that does not necessarily mean that the artefact is part of the cognitive process itself. The modified EM is immune to this fallacy because it specifies a more rigorous criterion for an external artefact to be considered as part of the cognitive process: the existence of a feedback loop involving active manipulation. This requirement goes beyond mere coupling and ensures that the artefact is not just associated with the cognitive process but is actively involved and essential to it. A feedback loop implies a dynamic, reciprocal interaction between the agent and the artefact. In such a loop, the agent actively manipulates the artefact, and the artefact, in turn, has a direct impact on the cognitive processes of the agent. The constitutive role is established through this active, reciprocal interaction, so that the cogniser and the artefact are mutually influential. This level of integration is limited to the specific context of the feedback loop. It is a targeted, task-specific interaction in which the artefact is a constitutive part of the cognitive process for the duration and purpose of that specific task.

Cognitive bloat is the expansion of the boundaries of the mind that occurs when too many external elements are considered as having a part in cognitive processes. In standard EM frameworks, this can lead to the problematic implication that almost any external object that an individual interacts with could be considered a part of their cognition. The modified EM avoids cognitive bloat by setting a clear criterion for what counts as a constitutive part of cognition: the presence of a feedback loop involving the manipulation of an external information-bearing structure. This criterion is specific and restrictive, and it prevents the indiscriminate inclusion of external artefacts in cognitive processes. Only those artefacts that are actively manipulated in a feedback loop – and thereby have a direct and significant impact on the cognitive task – are considered as being able to extend cognition. The nature of the interaction in the feedback loop is what grants the artefact its constitutive role. By defining specific criteria for cognitive extension – particularly the requirement for a feedback loop involving the manipulation of external artefacts, the modified EM effectively addresses the concerns of the coupling-constitution fallacy and cognitive bloat.

The following section evaluates various examples from the EM literature to determine whether they are suitable candidates for EM.

3.7.2 EM Examples

Otto

C&C formulate the example with Otto to establish the functional equivalence of Inga's biological memory and Otto's notebook. Since the modified version of EM is not based on functionalism, such examples have to be evaluated based on the role of feedback loops. In the new version of EM, the manipulation of an external information-bearing structure via a feedback loop is necessary and sufficient for EM. In Otto's example, the relation between Otto and his notebook cannot be considered as a feedback loop; therefore, according to this modified EM, this case cannot be considered as an example of EM. According to Palermos's criterion for EM, which requires a dynamic feedback loop with CRC, Otto's case cannot be considered as an example of EM because the connection between Otto and his notebook is intermittent. However, Palermos argues that, in many instances, normal memory appears to be an intermittent one-step process of storage and retrieval that is used when required. Otto's relation with his notebook is the same as that with his memory. Therefore, it was not clear for Palermos whether Otto's case is an example of EM. He proposes:

> In effect, the answer to the question whether, by the lights of the CRC criterion, Otto's distinct time (but still ongoing) interaction with his notebook counts as a case of cognitive extension depends on whether *distinct time dynamical systems* can produce the same coupling arguments [as] *continuous time dynamical systems.* (Palermos, 2014b, p. 38)

In the modified version of EM, a constitutive role for an external artefact in cognition requires a dynamic feedback loop involving active manipulation by the agent. This means that the agent not only uses the artefact but also actively changes or influences it as part of the cognitive process. Active manipulation implies a two-way interaction in which the artefact is not just passively used for information storage or retrieval but is actively altered or modified in a way that influences the cognitive task. Otto's notebook functions primarily as a storage device for information, which he retrieves when needed. Otto writes in the notebook and consults it, but this interaction lacks the dynamic, reciprocal quality emphasised in the modified EM. The notebook is not actively manipulated or altered as part of a feedback loop in the cognitive process of remembering; it serves as a static repository. Otto's interactions with the notebook are intermittent and do not involve a continuous feedback loop in which both Otto and the notebook are influencing each other in a dynamic way. The notebook's role is more akin to that of an enabling tool rather than being a constitutive part of the cognitive process; therefore, Otto's notebook is not a case of EM.

Ada

Clark (2008) describes an accountant, Ada, whose keen ability with numbers is the result, not from her making onerous demands on her biological memory, but from her "scanning the columns, copying some numbers onto a paper scratchpad, and then looking to and from those numbers (carefully arrayed on the page) back to the columns of figures". Clark, in reference to Ballard et al. (1997), describes Ada as employing a number of "minimal memory strategies":

> Instead of attempting to commit multiple complex numerical quantities and dependencies to biological short-term memory, *Ada creates and follows trails through the scribbled numbers*, relying on self-created external traces every time an intermediate result is obtained. These traces are visited and re-visited on *a just-in-time, need-to-know basis*, briefly shunting specific items of information into and out of short-term bio-memory in much the same way as a serial computer shifts information to and from the central registers in the course of carrying out some computation. (Clark, 2008, p. 69)

Thus, rather than solving an accounting problem in her head, Ada engages in a complex back-and-forth with her environment that involves "a distributed combination of biological memory, motor actions, external symbolic storage, and just-in-time perceptual access" (Clark, 2008, p. 69). Although Ada's strategies involve external artefacts (the paper scratchpad), the nature of her interactions with these artefacts may not constitute a reciprocal feedback loop as defined in the modified EM. Ada uses the scratchpad to offload cognitive tasks, so that it functions more as an enabling tool rather than a constitutive element in a dynamic cognitive process. The interaction lacks the continuous, two-way engagement characteristic of a feedback loop in which the external artefact is not just used but is actively altered or manipulated in a way that is integral to the cognitive process.

According to the new version of EM, a feedback loop with an external information-bearing structure is necessary and sufficient for EM. There is no reciprocal feedback loop in Ada's case. Ada does not manipulate an external artefact to accomplish epistemic tasks. Rather, Ada uses external artefacts to offload a cognitive task, which can be explained in terms of the external artefacts having an enabling role. Therefore, Ada is not an example of EM.

An artist

Clark (2001) claims that when an artist is interacting with their sketchpad, then this interaction can play a constitutive role in the cognitive processing by the artist. Clark argues that there is evidence to suggest that by externalising an image through drawing or sketching, an agent can manipulate and transform that image in ways that they could not do by internal means alone. Clark (2001) argues that the research by both Chambers and Reisberg and Van Leeuwen et al. demonstrates that: "Human thought is constrained, in mental imagery, in some very specific ways in which it is not constrained during online perception." This leads Clark to conclude that when we think of an artist and their sketchpad, we should recognise that:

> The use of the sketchpad is not just a convenience for the artist, nor simply a kind of external memory, or durable medium for the storage of particular ideas. Instead, the iterated process of externalising and re-perceiving is integral to the process of artistic creation itself. (Clark, 2001)

Clark's claim then is that since the interaction between the artist and their sketchpad is integral to the process of artistic creation (as follows from the research by Van Leeuwen et al.), then the interaction should be viewed as a constitutive part of the cognitive processing of the artist. In which case, the vehicles responsible for cognition by the artist will extend to include that interaction. If this is correct, then the cognitive processing of the artist will, under these particular circumstances, extend into the world.

Clark's argument that the artist's interaction with their sketchpad constitutes an essential part of their cognitive process aligns with the idea of it being an information-bearing structure in some respects. The artist externalises their thoughts through drawing, thereby manipulating and transforming the visual representation on the sketchpad. This process is iterative and dynamic with a continuous interaction between the artist and the sketchpad. However, under the modified EM framework, the sketchpad needs to be more than just a medium for externalisation. It should actively contribute to the cognitive process in a way that is integral and indispensable. Although the sketchpad is certainly a cognitive aid that allows for the externalisation and re-perception of ideas, it may not meet the stricter criterion of being an information-bearing structure that is actively manipulated as part of a reciprocal cognitive process. The sketchpad serves primarily as a medium for externalising and visualising thoughts, but it may not be actively manipulated in the sense required by the modified EM. The sketchpad does not undergo any transformation or active manipulation that feeds back into the cognitive process in the same way as, for example, a digital tool that adapts its output based on user interaction. If the sketchpad is seen as a passive recipient of the artist's output rather than an active participant in a feedback loop, then it is not a constitutive part of the cognitive process as defined in the modified EM.

Globe Theatre

Sutton (2010) endorses Tribble's (2005) study of Shakespearean actors at the Globe Theatre. Sutton argues that this study demonstrates how the environmental resources in the Globe acted as an external memory resource, since they had a crucial role in enabling actors to learn and memorise their cues and dialogue. In other words, these resources had transformative potential for the memory of these actors because they enabled the actors to complete cognitive tasks that would have proved impossible (or at least, extremely difficult) without such resources. Sutton considers that this is an example of EM.

According to the new version of EM, the manipulation of an external information-bearing structure via a feedback loop is necessary and sufficient for EM. Thus, the Globe Theatre cannot be considered as an example of EM as there is no interactive feedback loop. Rather, it has fixed external memory aids. The enabling role of such cognitive aids are sufficient to explain this example.

Night vision goggles and a soldier

Pritchard (2018a) explains what a candidate for extended knowledge would look like using an example of night vision goggles. When a soldier uses night vision goggles for the first time, the relation between the agent and the device is like that between a subject and an instrument. When the subject uses the instrument, a feedback loop is set up (e.g. what she sees will guide where she looks, so she learns to adjust the settings of the device to suit her preferences, and so on). However, over time, the soldier will become familiar with the instrument and will completely integrate it into her overall cognitive processes in a seamless fashion so that she forms beliefs unreflectively. Pritchard (2018a) considers the seamless cognitive processes in using night vision goggles as a candidate for extended knowledge.

I argued that a soldier using night vision goggles is an example of EM because night vision goggles are highly complex. They have patterns and histories embedded within them. Manipulating such tools via a feedback loop enables one to accomplish cognitive tasks that could not otherwise be fulfilled via normal open loops.

3.7.3 Conclusions

The modified version of EM is not based on functional equivalence or the parity principle. The modified version of EM requires:

- an external information-bearing structure that has associated patterns or a history
- an agent who establishes a feedback loop that results in the manipulation of the external information-bearing structure to accomplish a particular epistemic task

As per the criteria above, some of the examples purporting to represent an EM in the literature (such as Otto, Ada, an artist with a sketchpad and the Globe Theatre) cannot be considered as examples of EM. However, a soldier using night vision goggles can be considered as an example of EM.

By requiring a feedback loop with active manipulation, the modified EM ensures that only those artefacts that are truly integrated into and transformative of cognitive processes are considered as being able to extend cognition. Since a feedback loop with the manipulation of an external information-bearing structure is necessary and sufficient for EM, there is no cognitive bloat. The modified EM addresses the coupling-constitution fallacy. This fallacy arises when there is confusion between an agent merely coupling with an external artefact and that artefact being a constitutive part of a cognitive process. In the modified EM, the requirement for active manipulation within a feedback loop clearly delineates when an external artefact transitions from being a mere tool (coupling) to being an integral component of cognition (constitution).

3.8 Integrated EM

What are the advantages of establishing the constitutional role of an external artefact against an enabling role? The advantages of recognising a constitutional role for external artefacts, as opposed to merely an enabling role, are significant, particularly in addressing the limitations of cognitive internalism. Cognitive internalism suggests that all cognitive processes occur solely within the confines of the individual's mind. This traditional perspective has difficulties in fully accounting for the complexity of human cognition, particularly in the context of social interactions and cultural developments. Social cognition in humans often involves interactions with and a reliance on external artefacts and structures. These interactions are not merely auxiliary but play a fundamental role in shaping cognitive processes, especially in complex tasks and social contexts. Acknowledging a constitutional role for external artefacts in cognition means accepting that these artefacts do more than assist or enable cognitive tasks; they become integral parts of the cognitive process. This perspective aligns more closely with how humans engage with their environment and utilise tools and artefacts in cognitive tasks. By integrating external artefacts into our understanding of cognition, we can better explain the construction and evolution of social cognition in humans. This integration helps account for the ways in which cultural practices, technological innovations and social interactions shape and are shaped by cognitive processes. This perspective offers a more holistic view of cognition, one that encompasses both internal mental processes and the external, artefact-mediated interactions that are fundamental to human cognitive development and social functioning.

Giere and Moffat (2003, p. 308) note in their discussion of the scientific revolution of the 16th century:

> No 'new man' suddenly emerged sometime in the sixteenth century. … The idea that a more rational mind … emerged from darkness and chaos is too complicated a hypothesis" [Latour 1986, p. 1]. We agree completely. Appeals to cognitive architecture and capacities now studied in cognitive sciences are meant to explain how humans with normal human cognitive capacities manage to do modern science. One way, we suggest, is by constructing distributed cognitive systems that can be operated by humans possessing only the limited cognitive capacities they in fact possess.
>
> (Giere & Moffat, 2003, p. 308; reproduced from Carter et al., 2014, p. 96)

Carter et al. (2014) emphasise the role and importance of HEC for integrating seemingly different fields to gain better insights. Carter et al. indicate that there is a potential link between epistemology and the philosophy of science based on extended cognitive characters. Epistemology and the philosophy of science are intimately related. Whereas epistemology developed as an individualistic discipline, the philosophy of science is socially oriented. Carter et al. claim that:

> Science is primarily performed by individual scientists employing their hardware and software epistemic artifacts or by research teams operating within scientific labs that are uniquely tailored to fit their purposes. Accordingly, the concepts of extended cognitive characters and epistemic group agents could become very handy for a mainstream epistemological analysis of the scientific progress. (Carter et al., 2014, p. 96)

However, what is the relevance of EC if it occurs only when a cognitive agent establishes a feedback loop with external information-bearing structures that results in the manipulation of the external information-bearing structures to accomplish a cognitive task? In the big picture of various cognitive loops and processes, is it important to explore such a narrow scope for feedback loops with external information-bearing structures? Here, I am trying to integrate HEC with NCT and DST in developmental systems to determine the role and boundaries of EC.

As well as their support for feedback loops with external information-bearing structures, which are sufficient and necessary for EM, theories such as NCT, DST and developmental systems theory provide an integrated approach to understanding cognition and the boundaries of EM. This section focuses on the scope of the integrated approach and the limits of integrated EM.

3.8.1 Intersections of NCT, EC, CN, CNC and Virtue Epistemology

In the previous sections, I explained how niche construction by an organism can result in evolutionary pressure on its own. Niche construction can result in cultural inheritance, which includes artefacts and cumulative learning passed down the generations. For humans, CNC, which is indebted to NCT, and CN explain human cognitive development via cause-and-effect reasoning and the construction of cognitive aids and artefacts. Humans collectively became a superorganism through cognitive development via language, sociality, tools and the transfer of cumulative knowledge through the generations. This cognitive development allows humans to adapt easily to various environments, unlike the fixed adaptability of animals to a particular environment. Moreover, developmental systems theory, which I discussed in Section 3.4, describes the cognitive development of someone from infancy to adulthood via learning and cultural inheritance.

EC claims that cognition extends into the environment via cognitive artefacts. The application of DST to cognition and developmental systems theory can explain the nature of the engagement of a cognitive agent with the environment via cognitive artefacts and cognitive development. Here, I summarise the various aspects of a cognitive agent and the interactions of the agent with the environment based on NCT, CNC, CN, developmental systems theory and DST. I explore the possible consequences of EC for the above-mentioned theories. I attempt to discern broad patterns in the cognitive development of humans and the role and boundaries of EC.

I structure the remaining part of this section into three core ideas:

1. The manipulation of an external structure by an agent, as proposed in EC, is aligned with the principles of NCT, which describes how organisms modify their environments. This intersection shows how cognitive agents shape and are shaped by their environments, leading to the creation of cognitive niches (CN and CNC). These niches represent the cumulative effect of the feedback loops, wherein both the cognitive processes of the agent and the external structure evolve through cycles, which contribute to the development of more sophisticated cognitive strategies and environments. This ongoing interaction, characterised by reciprocal influence and adaptation, forms a key aspect of both EC and NCT and illustrates their interplay in the evolution of human cognition and environmental interaction.
2. The intersection of various theories contributing to an understanding of cognitive development.
3. An attempt to establish the core aspects and the broad patterns of cognitive development based on NCT, EC, DST and developmental systems theory.

3.8.1.1 NCT, NC, EC and Virtue Epistemology

As knowledge is an important aspect of cognition and the development of cognitive patterns, to understand the consequences of NCT, CNC and EC, I use terms from virtue epistemology (VE), such as cognitive character, cognitive ability and cognitive traits, to establish the relations among NCT, EC and VE. For VE, especially virtue reliabilism, the reliability of cognition is not just about internal cognitive faculties but also involves the external resources and environments that have been adapted by humans. This view is aligned with cognitive extension in EC and is supported by the evolutionary insights provided by NCT. The integration of externalist VE (specifically virtue reliabilism), EC and NCT forms a comprehensive framework that spans philosophy, epistemology and evolutionary biology. It offers a nuanced understanding of cognition as an adaptive, extended process that is deeply interconnected with the environments that humans construct and interact with. This cross-disciplinary approach provides a holistic view of human knowledge acquisition by considering both the internal cognitive faculties and the external resources and environments that shape and support them.

VE is a complete theory of knowledge in which knowledge is a cognitive achievement of an agent's beliefs formed by exercising their cognitive ability. In VE, only those reliable belief-forming processes that make up one's cognitive character – such as one's cognitive faculties, cognitive abilities and intellectual virtues – can generate knowledge (Pritchard, 2018b).

The externalist version of VE, i.e. virtue reliabilism, allows knowledge to accumulate through the appropriate functioning of an agent's cognitive abilities and faculties. Pritchard (2018a) explains that cognitive abilities are innate and passive and that knowledge is normally acquired via purely unreflective routes. Pritchard (2018b) claims that this is especially true of one's cognitive faculties. Pritchard further suggests that intellectual virtue is a motivational state, characteristic of a good inquirer. It is a sophisticated, acquired and active cognitive trait rather than a mere cognitive ability or faculty.

NCT is based on feedback loops with the environment at a genetic level and at an ontogenetic level. The ontogenetic and cultural feedback loops in NCT can explain the rapid development of humans compared with our nearest relatives in the evolutionary chain. Cognitive niche construction and ecological inheritance explain how humans interact with the environment and how humans develop their intelligence through technical know-how, tool use, grammatical language, social learning and cooperation and how the accumulated knowledge is transferred across generations.

CNC and CN can explain human cognitive development based on aspects of niche construction and also explain cause-and-effect improvisational intelligence. In the feedback loops and human cognitive development described by NCT, broad patterns can be visualised based on the development of individual psyches via social learning, cultural inheritance and cognitive practices across society.

The broad patterns derived from NCT and EC in relation to VE are:

1. Cultural inheritance, cognitive practices and cognitive patterns across society: These can be compared with the intellectual virtue of VE, i.e. acquired, habituated and active cognitive traits. In EC, such cognitive traits and practices can cause feedforward loops, such as the emulation model discussed in Section 3.2.

2. New developments in abstract reasoning through inventions and new concepts: These can be compared with the results of the acquired, active and motivational state of intellectual virtue. The highest outcome of intellectual virtue is the identification of new patterns.

3. The development of patterns of cognitive and cultural practices by a population through learning about new developments in abstract reasoning via psycholinguistic metaphorical abstraction: This can be compared with the initial set-up of feedback loops between an agent and an external artefact and the manipulation of the external artefact in EC, as in the example of night vision goggles operated by a soldier. Such a feedback loop can be compared to a cognitive ability in VE.

4. Genetic features of NCT and feedback loops: These are comparable to the innate cognitive faculties and cognitive abilities in VE.

NCT emphasises the ongoing interaction between organisms and their environment. Organisms actively modify their surroundings, which in turn influences their development. This concept is aligned with EC, which postulates that cognitive processes can extend beyond the brain to include external artefacts and environments. VE, specifically its focus on intellectual virtues, adds another layer to this interaction. Intellectual virtues are sophisticated cognitive traits that involve a motivational state characteristic of good inquiry. They are not just passive faculties but actively acquired and refined through interaction with the environment.

The intersections of NCT, EC and VE are shown Figure 3.16. There are dynamic interactions between each domain of the triangle on the right-hand side of the figure. Each domain is interconnected with every other domain.

The feedback loop between an agent and an external artefact, such as in the example of night vision goggles and a soldier, is shown at the bottom of Figure 3.16. Once the solider is familiar with the operation of the night vision goggles, it becomes a cognitive ability or an extended cognitive faculty. Thus, such a feedback loop enables us to acquire knowledge.

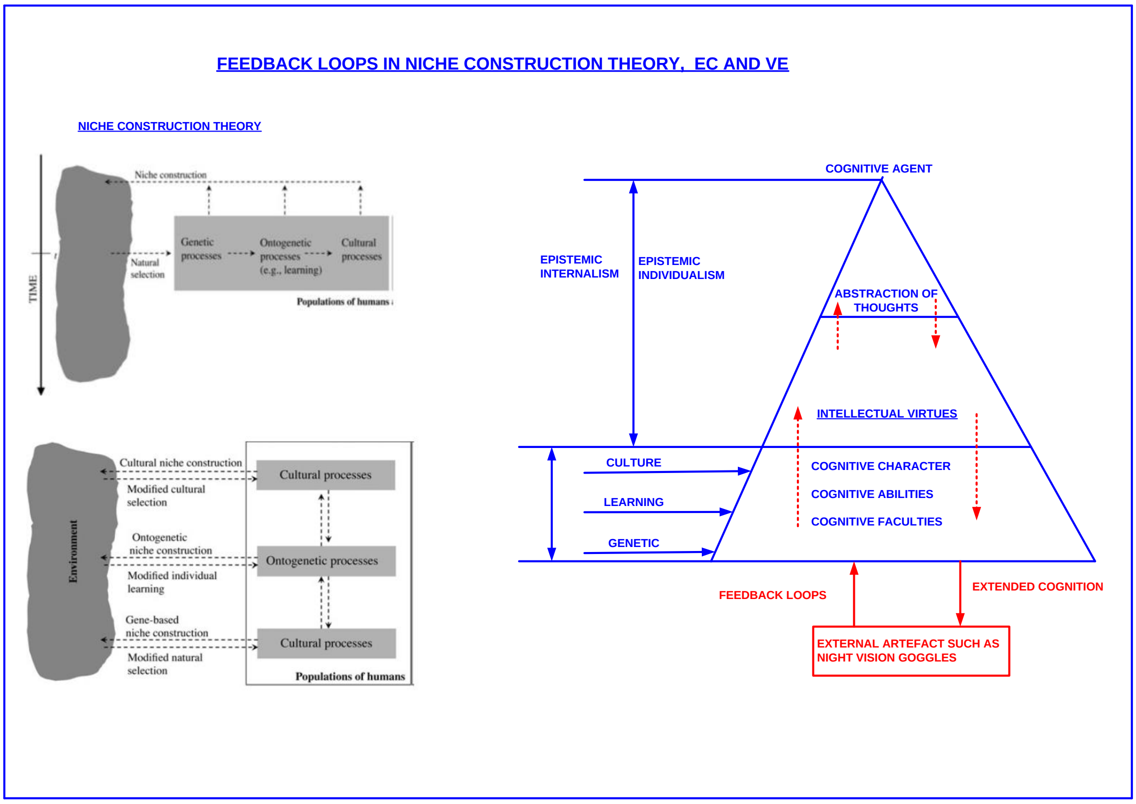


Figure 3.16. Intersection of NCT, EC and VE.

The cultural feedback loops in EC, through which organisms interact with and are influenced by their social and cultural environments, also intersect with NCT. These interactions contribute to the formation of beliefs and cognitive processes. Beliefs formed through cognitive processes that involve external artefacts, as in EC, are influenced by the environmental interactions highlighted in NCT. This influence extends to the development of intellectual virtues, as posited in VE. The intersection of NCT, EC and VE presents a comprehensive view of cognition that encompasses genetic, ontogenetic and cultural dimensions. It underlines the dynamic nature of cognitive development, which is influenced by both internal abilities and external interactions. This integrated perspective helps explain how cognitive traits and abilities evolve, not just through internal mechanisms but also through active engagement with and adaptation to the external world. The interplay between NCT, EC and VE provide a holistic understanding of cognition and highlights the importance of external interactions and cultural influences in the development of cognitive faculties and intellectual virtues. This approach underscores the dynamic and adaptive nature of human cognition, as shaped by both the environment and internal capacities.

Three domains can be visualised from the interaction of a cognitive artefact with an agent. Each domain encompasses the others. The bottom domain is the feedback loop between the agent and the artefact. The second domain is the acquired abilities and intellectual virtues that arise from the innate cognitive abilities of the agent through their interaction with the artefact. The upper domain is abstract thought, which originates inventions and new ideas. There are dynamic interactions between each domain, and each is interconnected with every other domain. As shown in Figure 3.16, NCT illustrates the feedback loops of an organism with the environment at the genetic and ontogenetic levels. The triangle on the right-hand side shows the potential intersection of NCT and VE. Based on NCT, Pritchard suggests that intellectual virtue is a motivational state and that cognitive traits are in the domain of the ontogenetic feedback loops between an organism and the environment. The sketch indicates that there is a dynamic interaction between each domain of the triangle. Moreover, each domain is interconnected with every other domain.

NCT demonstrates how organisms engage in feedback loops with their environment at both the genetic and ontogenetic levels, as they are shaping and being shaped by their surroundings. This concept is aligned with the VE domain, in which the development of intellectual virtues can be seen as part of these ontogenetic feedback loops. These virtues emerge as organisms interact with and adapt to their environment, thus reflecting the dynamic nature of cognitive development. Additionally, the beliefs formed through cognitive processes in which an external artefact plays a constitutive role are situated within the NCT domain. These beliefs are products of ontogenetic feedback loops between organisms and their environments, which illustrates the influence of external factors on cognitive development. Similarly, the cultural feedback loops described in EC, which encompasses interactions between organisms and their socio-cultural environments, are also aligned with the principles of NCT.

Cognitive abilities are the fundamental mental skills we use to think, learn and understand. When these abilities are refined and directed towards productive inquiry, they can develop into intellectual virtues. Such virtues are active and motivational states that drive individuals to engage deeply with ideas and problems. These virtues push individuals to go beyond basic understanding to more complex and abstract thinking. Abstract reasoning is the ability to understand complex concepts that are not grounded in physical reality or immediate experiences. It involves thinking about ideas, patterns and principles that are abstract or theoretical. The active and motivated state of an intellectual virtue promotes the development of abstract reasoning. When someone is intellectually curious and open-minded, they are more likely to explore complex ideas and identify new patterns in information. One of the outcomes of enhanced abstract reasoning, which is fuelled by intellectual virtues, is the ability to identify new patterns. This means recognising connections or principles that were not apparent before. This ability to see new patterns can lead to inventions or novel solutions to problems. It is the process of taking abstract ideas and turning them into something concrete or useful. Over time, this practice of abstract reasoning and pattern recognition becomes integrated into our cognitive framework. An abstract idea that was once a novel concept gradually transforms into a familiar cognitive pattern or even a refined intellectual virtue. This transformation leads to an enhancement of our overall cognitive abilities. As our cognitive abilities are enhanced, they further strengthen our intellectual virtues. This enhancement creates a positive feedback loop: improved intellectual virtues lead to more sophisticated abstract reasoning, which in turn leads to the further development of intellectual virtues. Intellectual virtues act as catalysts for developing abstract reasoning and identifying new patterns. This process is not just a one-time event but a continuous cycle in which enhanced cognitive abilities lead to the further development of intellectual virtues, thus fostering a cycle of cognitive growth and the generation of new ideas. The cycle continues: enhanced cognitive abilities lead to higher-level intellectual virtues, which further foster the development of advanced abstract thoughts. This ongoing process can be thought of as a spiralling rise in cognitive and intellectual development.

In essence, the evolution from basic cognitive abilities to advanced intellectual virtues, and from there to innovative abstract thinking, is a dynamic and iterative process. Thus, intellectual growth is not a linear but a progressive development in which each stage builds upon and enhances the previous one. The identification of new patterns or inventions is a critical milestone that marks the point at which abstract thinking, guided by intellectual virtues, materialises into concrete outcomes. Thus, the interplay between cognitive abilities, intellectual virtues and the use of external artefacts is not static but dynamic and cyclical. It is a continuous process of growth and enhancement, such that each stage of cognitive development feeds into and elevates the next, leading to an ever-expanding realm of intellectual capability and innovation.

The above discussion shows that the dynamic process through which cognitive abilities, when engaged with external artefacts, can develop into intellectual virtues and lead to advanced abstract thinking and the continuous enhancement of cognitive capacities.

Figure 3.17 illustrates the establishment of a feedback loop between an agent and an external artefact and the development of that feedback loop to become an abstract thought, which then evolves into an open loop and feedforward loop at various times, as in the emulation model. It clearly separates the feedback loop and the constitutive role of the external artefact from its enabling role. At the boundary, the cognitive ability changes to intellectual virtue and abstraction of thought.

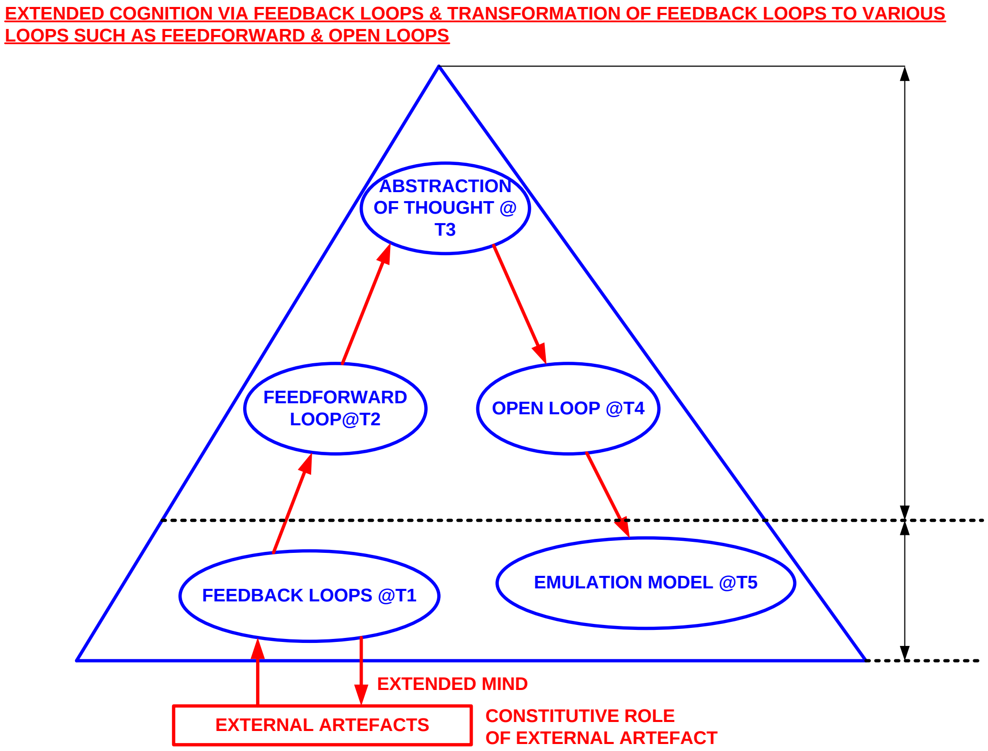

EXTENDED COGNITION VIA FEEDBACK LOOPS & TRANSFORMATION OF FEEDBACK LOOPS TO VARIOUS LOOPS SUCH AS FEEDFORWARD & OPEN LOOPS
ABSTRACTION OF THOUGHT @ T3
FEEDFORWARD LOOP@T2
OPEN LOOP @T4
FEEDBACK LOOPS @T1
EMULATION MODEL @T5
EXTENDED MIND
EXTERNAL ARTEFACTS
CONSTITUTIVE ROLE OF EXTERNAL ARTEFACT

Figure 3.17. Extended cognition via feedback loops and the transformation of feedback loops to various types of loop, such as feedforward and open loops.

Once a feedback loop has become established as part of an agent's cognitive character, it can be seen as a cognitive ability and can give rise to open or feedforward loops. Such a transformation of an acquired skill makes EC more complex. My argument is that in EC, a feedback loop with the external artefact and the manipulation of that artefact are necessary and sufficient. However, with dynamic progress, such a feedback loop can become the basis for an acquired skill or a developed cognitive pattern, which itself can then become the basis for further open and feedforward loops during cognition.

Figure 3.18 illustrates the broad patterns in developmental systems theory along with NCT, CN, EC and VE. It is complex to picture the cognitive development of humans from the perspective of genetics, the cultural inheritance aspects of evolution, CN, EC and developmental systems theory.

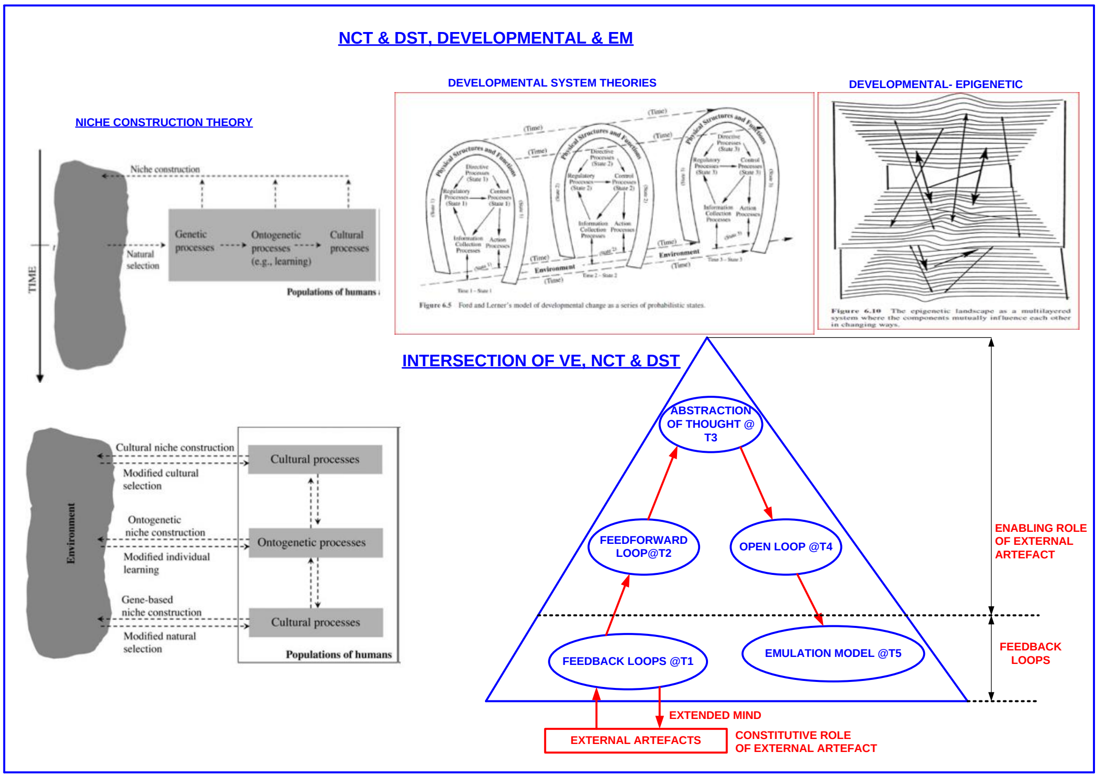

NCT & DST, DEVELOPMENTAL & EM
NICHE CONSTRUCTION THEORY
TIME
t
Niche construction
Natural selection
Genetic processes
Ontogenetic processes (e.g., learning)
Cultural processes
Populations of humans
DEVELOPMENTAL SYSTEM THEORIES
Physical Structures and Functions
Directive Processes (State 1)
Regulatory Processes (State 1)
Control Processes (State 1)
Information Collection Processes
Action Processes
(Time)
Environment
Time 1 - State 1
Time 2 - State 2
Time 3 - State 3
Figure 6.5 Ford and Lerner's model of developmental change as a series of probabilistic states.
DEVELOPMENTAL- EPIGENETIC
Figure 6.10 The epigenetic landscape as a multilayered system where the components mutually influence each other in changing ways.
INTERSECTION OF VE, NCT & DST
Environment
Cultural niche construction
Modified cultural selection
Ontogenetic niche construction
Modified individual learning
Gene-based niche construction
Modified natural selection
Cultural processes
Ontogenetic processes
Cultural processes
Populations of humans
ABSTRACTION OF THOUGHT @ T3
FEEDFORWARD LOOP@T2
OPEN LOOP @T4
FEEDBACK LOOPS @T1
EMULATION MODEL @T5
ENABLING ROLE OF EXTERNAL ARTEFACT
FEEDBACK LOOPS
EXTENDED MIND
EXTERNAL ARTEFACTS
CONSTITUTIVE ROLE OF EXTERNAL ARTEFACT

Figure 3.18. Intersection of NCT, DST, developmental systems theory and EM.

From the NCT feedback loops and human development, broad patterns can be discerned based on the development of individual psyches via social learning, cultural inheritance and cognitive practices across society.

Pinker (2010) states that CN can explain the evolution of advanced language and intelligence. However, since the cognitive mechanisms were selected for physical and social learning, how could they have enabled *Homo sapiens* to engage in the highly abstract reasoning required for modern science, philosophy, government, commerce and law? Pinker (2010) argues that at most times, places and stages of development, most people are not readily using highly abstract reasoning; however, all of us are capable of learning about the inventions and abstract ideas developed by others. This capacity for learning is connected with a psycholinguistic phenomenon that may be called metaphorical abstraction, which reflects the ability of the human mind to readily connect abstract ideas with concrete scenarios. Pinker (2010) believes that:

> Humans possess an ability of metaphorical abstraction, which allows them to co-opt faculties that originally evolved for physical problem-solving and social coordination, apply them to abstract subject matter, and combine them productively. These abilities can help explain the emergence of abstract cognition without supernatural or exotic evolutionary forces and are in principle testable by analyses of statistical signs of selection in the human genome. (Pinker, 2010, p. 8993)

New inventions, new concepts and progress in the abstraction of thought, which are developed by only a few humans, can be learned by others through this psycholinguistic metaphorical abstraction. This results in a feedback loop of patterns, cognitive practices and social learning. Pinker (2010) states that humans have the following unique hyper-developed features compared to the rest of the animal kingdom: (1) cooperation among non-kin, (2) tool use and technological know-how and (3) grammatical language. As a result, science and abstract thinking have emerged.

3.8.2 Core Aspects of Cognitive Development

The intersections of NCT, EC, CNC and VE provide a theoretical foundation for understanding how humans interact with their environment and how this interaction shapes cognitive processes. The next step for us is to explore how these theories are manifest in concrete aspects of human cognition. This section transitions the discussion to a focused exploration of the core aspects of human cognitive development.

The following sections provide insights into the core aspects of human cognitive development, such as tool use, technical know-how, language, epistemic engineering and sociality. Each of these areas is a crucial facet of cognitive development and can be used to illustrate the practical applications and implications of the theories previously discussed. For instance, tool use and technical know-how will be examined through the lens of EC and NCT to show how cognitive processes extend and evolve through their interaction with our environment. Similarly, the exploration of language and sociality will use the principles of VE and CNC to highlight the social and cultural dimensions of cognitive development. This section bridges the gap between the abstract theoretical frameworks of NCT, EC, CNC and VE and their tangible expressions in everyday cognitive activities. By examining these core aspects, I explore a deeper and more comprehensive understanding of the dynamic interplay between an individual's cognitive capabilities and the external world as well as the social and cultural factors that shape and define human cognition.

3.8.2.1 Tool Use and Technical Know-how

From the previous discussions it is evident that tool use and technical know-how serve a key role in the cognitive development of humans. Pinker (2010) argues that the complex tools used by humans are deployed in extended sequences of behaviour, which are acquired both by individual discovery and by learning from others. Using tools is important for human survival and higher-order mental functions.

Vygotsky (1978) suggests that a tool is a mediating link between the actions of a person and an object. There are technical and psychological tools. Technical tools are used with the intention of creating changes in the external physical world whereas psychological tools act upon mind and behaviour, facilitating activity towards oneself and not towards an object. For Vygotsky, number systems, mnemonic techniques, diagrams and maps are psychological tools. Vygotsky claims that using psychological tools enhances and immensely extends the possibilities of human behaviour. We develop our cognitive capacity through the creation and manipulation of external resources, such as signs, diagrams and maps. Vygotsky (1930) states that: "In the behaviour of men we encounter quite a number of artificial devices for mastering his own mental processes. … These devices can justifiably and conventionally be called psychological tools or instruments" (Vygotsky, 1930, p. 1).

According to Clark (1997a), in finding efficient and systematic ways of fulfilling our goals (and exploiting information-bearing structures in the environment), humanity has developed tools and even designed environments that give us ever greater abilities and allow us to achieve goals that would otherwise be beyond us. Such tools utilise the kinds of inner reasoning and outer manipulation that fit our brains, our bodies and our evolutionary heritage. Our visual acuity and pattern-matching skills, for example, far outweigh our capacity to perform complex arithmetical operations (Clark, 1999). Regarding tool use and brain plasticity, the research that Clark cites demonstrates cases where a tool user's "plastic neural resources become recalibrated in the context of goal-directed whole agent activity" (Clark, 2008, p. 39).

3.8.2.2 Language

Combinatorial language plays key role in the cognitive development of humans. Clark (2005) argues that language occupies a cognitive niche and has three distinct but interconnecting roles:

1. As an augmented reality overlay in which the simple act of labelling reduces the computational burden and functions by open-endedly projecting new groupings and structures on to a perceived scene

2. As a scaffolding action, such that a language-using agent can engage in behavioural self-scaffolding, such as memorising instructions

3. As self-knowledge and mind control, which shows the human capacity to use linguistic rehearsal as a means of directing our own thoughts and reasoning

He argues that for thought, language is a self-constructed behaviour-enhancing niche, a super niche. Pinker (2010) claim that humans alone can use open-ended combinatorial grammatical language, which gives advantages in coordination, communication and the transmission of information.

Combinatorial grammatical language has a key role in social niche construction and cooperation among people. It allows an individual to interact with society and with the environment using social affordance, technical know-how and tool use. Cognitive niche construction, social niche construction and their interaction with the environment demonstrate how social learning enables cognitive practices, since the majority of people can learn about and assimilate abstract concepts.

3.8.2.3 Epistemic Engineering

Epistemic engineering plays a key role in the cognitive development of humans. Humans are epistemic engineers, i.e. humans make artefacts to gain knowledge and unveil the patterns. As Sterelny (2012) claims, humans construct epistemic niche of the next generation. Wheeler and Clark (2008) argue that: "Cognitive niche construction shows the actively engineered epistemic resources in the evolution and development of human cognition" (Wheeler & Clark, 2008, p. 3565).

Sterelny (2001) claims that organisms engineer their own environment, which is sometimes patterned across generations. Ecological inheritance supplements genetic inheritance. The uniqueness of humans lies in the exceptional capacity of human ecological engineers and active constructors of our cumulative and epistemic enriching cognitive niche, which is transferred to subsequent generations. Sterelny claims that humans are ecological engineers and that we are the active constructors of our own cognitive niche. Unlike other niche-construction animals, only humans accumulate downstream epistemic engineering, which means that knowledge and epistemic tools accumulate down the generations. Sterelny (2004) argues that: “Ecological engineering is visible to selection, for such alterations often have fitness effects that are stable across generations and niche-constructing behaviour itself evolves” (Sterelny, 2004, p. 231). Epistemic artefacts are tools for thinking, and they are central to the explanation of human intelligence and reasoning.

Sterelny (2003) states that our special cognitive powers derive from our ability to extend our minds’ capacities through interacting with our environment. Some of these ability-enhancing interactions are:

1. Our use of epistemic tools that make memorising easier
2. Our ability to transform difficult cognitive problems into simple sketches
3. Our ability to change difficult perceptual problems into easier ones using tools, for example, shaping wood with a chisel and hammer
4. Our ability to make difficult learning problems easier and to change the informational environment for the next generation
5. Our ability to engineer our workspace so that we can carry out tasks rapidly and reliably

He states that the evolution, operation and development of human intelligence are based on our ability “as epistemic agents, transforming the informational load on our own decision-making and that of others, and the ways we act as epistemic tool makers, constructing devices that help us meet those transformed informational challenges” (Sterelny, 2004, p. 253).

3.8.2.4 Sociality and Social Niche Construction

Sociality and cooperation are two important aspects of human cognitive development. C&C argue that there is a potential for socially EC, such as: "The waiter at my favourite restaurant might act as a repository of my beliefs about my favourite meals. … In other cases one's beliefs might naturally be seen to be embodied in one's secretary, one's accountant or one's collaborator." C&C claim that because of the extended self, "interfering with someone's environment will have the same moral significance as interfering with their person" and that "certain forms of social activity might be reconceived as less akin to communication and action and more akin to thought" (C&C, 1998, p. 18).

Sterelny argues that the evolution of accumulated social learning was a central causal factor in the evolution of human uniqueness. The uniqueness of the human mind can be explained in terms of cognitive niche construction and cultural inheritance. The model explains human uniqueness in terms of phenotypical and developmental plasticity, which is an adaption to the variability of our environments (Sterelny, 2012).

The social origin of mind describes how the cognitive processes normally associated with an individual mind can occur in a group of individuals. Vygotsky (1978) argues that every high-level cognitive function appears twice: first as an inter-psychological process and later as an intra-psychological process. Vygotsky argues that a child's inner speech has its origin in social contact. He states that in our conception, the true direction of the development of thinking is not from the individual to the social but from the social to the individual (Vygotsky, 1934).

3.8.3 Broad patterns of Cognitive Development: An Integrated Approach

Core aspects of cognitive development were analysed in the previous section by considering the consequences of NCT, EC, DST and developmental systems theory. Thus, broad patterns of cognitive development can be discerned.

The broad patterns, such as extended emotions, the reality template and abstract thought, collectively represent an integrated approach for analysing key aspects of cognitive development. They underline the complex interplay between internal cognitive processes and external environmental factors in shaping human cognition. This approach offers a comprehensive view of cognitive development, as it acknowledges the importance of both internal mental faculties and external influences in the formation of cognitive abilities and perspectives. The integrated approach can be summarised as domains of:

1. Emotions and extended emotions: This domain recognises that emotions are fundamental to human cognition. They are not just internal experiences but are often extended through interactions with others and the environment. For instance, shared experiences can lead to collective emotional responses, which demonstrates how emotions can extend beyond an individual's boundaries. Extended emotions also encompass how our feelings are influenced and shaped by cultural norms, social interactions and even the use of technology. This highlights the interconnectedness of emotional experiences with external factors.

2. Reality template: This is the sum and total of the cognitive patterns of an agent derived from cultural inheritance, affordance, tool use and various epistemic loops with the external environment at a particular time. It is a form of dynamic equilibrium. A reality template is an individual's comprehensive cognitive framework at any given time. It is a union of the cognitive patterns formed through cultural inheritance, affordance recognition, tool use and interactions with the external environment. A reality template is not static but a dynamic equilibrium that continually evolves as the individual interacts with and adapts to their environment. It includes learned patterns, perceptions, beliefs and skills that guide how an individual interprets and responds to their world. A reality template represents how an individual's cognition is shaped by both internal mental processes and external influences, including social and cultural factors. It captures the integrated and dynamic nature of cognitive development.

3. The abstraction of thought to identify new patterns: This domain is the cognitive process of abstract thinking, which is essential for identifying new patterns, solving problems and generating innovative ideas. It involves moving beyond concrete, immediate experiences to conceptualise broader principles or connections. The ability to think abstractly and identify new patterns is a crucial aspect of cognitive development. It signifies advanced cognitive functioning and is influenced by a combination of innate abilities, learned skills and interactions with the external environment.

Figure 3.19 shows the domains. Each domain progressively encompasses the underlying layers, and the domains are interconnected.

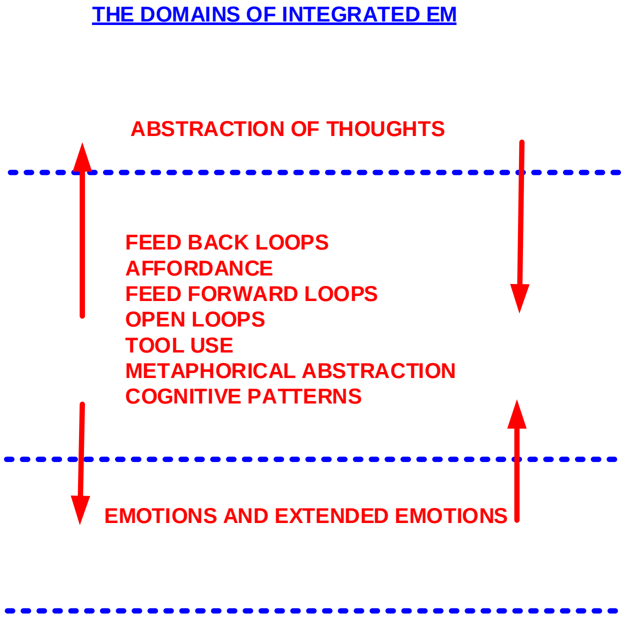


Figure 3.19. Domains of integrated EM.

3.8.3.1 Emotions and Extended Emotions

Theories of embodied emotions account for the bodily role in emotions, and they propose that emotions can extend beyond the brain. Damasio (1994) considers that emotions have a role in rational decision-making.

Traditionally, emotions and rationality are considered as distinct, so that emotions are considered to be irrational. However, Solomon (2003) and Nussbaum (2001) propose cognitive accounts of emotions. Solomon (2003) argues that: “Emotions require rationality (the ability to manipulate concepts) but they may be said to be rational or *ir*rational (opposed to *non*-rational) in a second sense, according to whether they succeed or fail to satisfy certain purposes or functions” (Solomon, 2003, Chapter 1, p. 22). Carter et al. (2016) state that a psychological account of emotions can be found in Aristotle, Hume, Descartes, Spinoza and Hobbes in which emotions are involved in some appraisals or evaluations. Carter et al. (2016) state that in a strong cognitivist picture, “indignance” is just a matter of judging, appraising or evaluating a certain behaviour as unfair, whereas “shame” is just a matter of judging oneself to have failed to have lived up to a particular ideal.

By combining the hypotheses of EC and EM with a dynamic interpretation of psychological appraisal theory, Carter et al. (2016) propose the extended emotion hypothesis. They argue that the emotions of an individual can extend into the environment and that they can be transferred to another individual using that person’s own cognitive processes without requiring a third entity, such as a group or collective mind.

Emotions are linked with patterns for survival. Whenever any action or event runs contrary to patterns of survival, negative emotions arise, such as anger, disgust, rage or jealousy. Whenever any action or event enriches the patterns of life, positive emotions occur. A negative emotion either results in a modification to the patterns for survival or completely changes them. Either way, the arousal of a positive or negative emotion is very much related to the patterns for survival.

The impact of the arousal of emotions varies for each individual. It depends on their patterns for survival and very much depends on the affordance and, in turn, on the position of the individual in society in terms of wealth, education and social recognition. However, the basic emotions of survival are common to all human beings.

3.8.3.2 Development of Cognitive Patterns via Metaphorical Abstraction and Reality Templates

I agree with Pinker (2010) that only a few humans invented the different components of modern knowledge, as we can we see from the history of scientific and technological progress. However, humans are capable of learning about the inventions and capable of abstract reasoning by metaphorical abstraction.

There are two aspects here:

1. Inventions (i.e. the capacity of some people to invent or identify new patterns through the abstraction of concepts)
2. The capacity of humans to learn and assimilate through such an abstraction of concepts. This is comparable to VE’s acquired, habituated and active cognitive traits.

Abstract concepts are assimilated by a population through the work of educated individuals via metaphorical abstraction, which again creates a feedback loop and results in patterns and cultural practices across society. This is comparable to VE’s acquired intellectual virtue.

Throughout history, we can identify the progress in learning, the assimilation of inventions and the abstraction of concepts by humans via psycholinguistic metaphorical abstraction. When a new invention or an abstract concept is learned by the majority of people, it becomes a cognitive pattern across society. Each human being develops their own patterns of life for survival. These patterns enable human beings to function in particular situations. Everyday life is based on these concepts or patterns of survival, which determine the overall functions or activities of the individual.

To understand the broad patterns in cognitive development, I propose reality templates. This concept captures the evolving nature of cognitive development over time and is a bridge that connects various interdisciplinary concepts within an integrated epistemological approach. A reality template has the following properties:

A dynamic equilibrium point: Much like the lowest point reached by a dampened pendulum, a reality template represents a state of balance in an individual's or society's cognitive framework. It is a point of stability within the constant flux of information and experience.

An attractor in cognitive dynamics: A reality template functions like an attractor in the dynamics of cognition. It is a point towards which cognitive processes tend. It shapes the agent's interaction with the environment and leads to the formation of new goals and understandings.

Regulated by various loops: The shaping of a reality template is influenced by different types of cognitive loops. These include feedback loops, where deviations from existing beliefs or understandings are detected and corrected; feedforward loops, which anticipate and prepare for future events; and open loops, which allow for the intake of new information without immediate feedback.

Formation and evolution of beliefs: A reality template is essentially the pattern of beliefs and knowledge that an individual holds at a given time. These beliefs are not static; they evolve based on interactions with the environment, occurrent beliefs, actions and offline reasoning. When new information is assimilated or a transient thought fades, the reality template adjusts to reflect a new equilibrium in the individual's cognitive state.

Micro- and macro-level templates: A reality template can exist at the micro level, specific to an individual, or at the macro level, encompassing societal beliefs. For instance, the societal understanding of space–time has evolved over the years, which illustrates how macro-level reality templates can shift with new scientific discoveries and theoretical advances.

Incorporating time and dynamism: The concept of a reality template introduces elements of time and dynamism into our understanding of cognitive development. It acknowledges that our cognitive frameworks are not fixed but are continually evolving in response to new experiences and information.

In summary, a reality template is a foundational concept that helps us understand how cognitive patterns develop and change over time. It accounts for the dynamic nature of cognition and highlights how our beliefs and understanding are continually shaped by our interactions with the world and evolve to reflect new knowledge and experiences.

Several examples of changes to reality templates can be found in physics, such as the changed conception of space and time. Newtonian physics considered that, in the laws of motion and gravity, space and time were absolute. However, in the special and general theories of relativity, Albert Einstein argues that space and time are relative. Einstein asserts that there are four dimensions instead of three, and he changes the concepts of space and time to space–time, thus explaining time dilation. The curvature of space–time can account for, for example, gravity, singularities and worm holes. Moreover, quantum mechanics uses a different reality template, based on uncertainty principles, measurement problems and observation issues, as well as quantum weirdness and paradoxes.

The above examples demonstrate how cognitive patterns form across society. When Einstein originally put forward the concept of relativity, few people could understand this abstract idea of space–time. As Stephen Hawking points out, millions can now understand the theory of relativity. This shows how abstract concepts can become transformed into a reality template, i.e. a cognitive pattern across society via psycholinguistic metaphoric abstraction.

3.8.3.3 Abstraction of Thought

The abstraction of thought or advanced conceptual abilities are an aspect of cognition. In modern cognitive science, these advanced conceptual capabilities are attributed to offline cognition. Wilson (2002) suggests that offline cognition might include planning, daydreaming and remembering. Clark and Grush (1999) claim that truly cognitive phenomena are those that involve offline reasoning, which is derived from online mechanisms. Online cognition is attributed to an immediate interaction with the world. As Wheeler (2005) argues, online cognition is a “flexible real-time adaptive response to incoming sensory stimuli”. Wheeler (2005) thinks that abstraction, as a form of cognition, is typically considered to be located within “offline cognition”.

Like the continuity we have discussed about the emergence of artefacts, throughout history, there is a line of progression in the development of abstract thinking, i.e. there has been progress in identifying patterns, formulating new concepts and developing inventions. The progress in identifying new patterns enables us to become closer to reality. Humans are aware of the continuity between history and the future, even though the lifespan of each individual is short. The development of the concept space and time is an example of progress in abstract reasoning.

Space is everywhere. This absolute space is where we exist. Put like that, it is simple and easy to comprehend; we can say that it is a common-sense concept. The ancients believed in absolute space, one that is aligned with our common-sense experience. When they realised that the Earth is spherical, that day and night occur because the Earth is spinning around its axis and that the seasons occur because of the rotation of the Earth around the Sun, this was an abstraction of the concepts of space and time, even though these ideas are now considered fundamental. When Newton discovered the laws of motion, which can predict the motion of objects, our conception of space and time changed again. Time became strongly related to motion. However, Newtonian laws did not contradict common-sense experience and were based on absolute time and absolute space, which our everyday knowledge is also founded on.

We retained this common-sense conception of space and time until Einstein’s theory of relativity. The experimental results of Michelson and Morley proved that the speed of light in a vacuum is a constant for all observers. This provoked Einstein to think deeply about space and time. To explain how the speed of light can be a constant with respect to each observer, Einstein considers that for different observers, time was relative. Thus, in relativity, time varies depending on the relative speeds of the observers. If two people each carry a clock and are travelling at different speeds, they will have different time frames, even though both clocks tick once every second. If one of the people is moving near to the speed of light, 1 second on their clock will last for years for a person walking at a speed of 1 m/s. The duration of time measured by the first clock therefore differs with respect to the experience of someone walking at a speed of 1 m/s. Hence, Einstein refutes the concept of absolute space and time, the one that is normal in common-sense applications. This conception of relative time is hard to comprehend, as it denies our common-sense experience of space and time.

So, Einstein links time with the relative motion of objects to explain how the speed of light is a constant, irrespective of an observer's location or motion. Subsequently, Einstein went on to prove that the force of gravity was due to the curvature of space–time. In this way, the abstraction of concepts has progressed in science, from Euclidean geometry to non-Euclidean geometry and from Newtonian gravity to Einstein's general theory of relativity.

Even though Einstein's conception of space–time deviates from our common-sense experience, it is still in line with how we consider that the observable world behaves in a deterministic way, as it does in Newtonian mechanics. Determinism suggests that we can predict the movement of objects in the world.

In contrast, the abstract concepts of quantum mechanics deny even the deterministic nature of the world. Quantum mechanics postulates the existence of probabilistic wave functions. Thus, an electron can have a probability of being in many different places at the same time. It has a probability for spinning both clockwise and anticlockwise simultaneously. Entanglement further stretches our concepts of space–time, since if two particles are entangled quantum mechanically, then making a measurement of a property of one particle instantaneously changes the same property of the other particle, irrespective of how far apart they are, which Einstein refers to as "spooky action at a distance". The cornerstone of quantum paradoxes lies in the contradiction between the space–time of quantum theory and a purely intuitive understanding of space and time based on common-sense experience.

Throughout history, there has been continued progression in the abstraction of thought processes, for example, from Newtonian physics to relativity and to quantum mechanics and the standard model. In summary, the argument that the abstraction of scientific thought processes is a result of the interaction between external and internal factors is grounded in the collaborative nature of scientific progress, the role of technological tools in extending cognitive capabilities, the reliance on empirical data and observation, and the continuous feedback loops between theory and experimentation. This perspective highlights that the evolution of complex scientific ideas is not merely a product of individual cognitive prowess but a complex process involving both the mind and its interaction with the external world.

3.8.3.4 Dynamic Development of Individual Psyches and Society

Reality templates can explain aspects of society too. For social cognitive development, a reality template encompasses cognitive patterns across society, such that at a particular time, the members of the society consider certain ideas to be true. For example, in ancient times, the notion of a flat Earth was a social reality template, as most people believed that the Earth was flat. However, after abstract reasoning, the notion of a spherical Earth replaced the reality template of a flat Earth.

Cognitive practices are patterns across society. A cognitive process used by an individual can extend into the environment and it can become transferred to another individual to become part of that person’s own cognitive processes without requiring a group or collective mind. As Sterelny (2001) argues, social learning is unique to humans.

According to socially extended cognition, the cognitive processes of an agent may spill from them to other cognitive agents, without a third entity being involved. In distributed cognition, cognitive agents work together to achieve joint cognitive success. In other words, cognition is distributed among several people who each participate in the cognitive task, such as each individual scientist in a large scientific project. Both EC and distributed cognition consider that social cognitive development occurs through the use of language and tools and through cooperation. Based on the cognitive nature of the agent, as illustrated in Figure 3.20, and based on EC and distributed cognition, similar kinds of cognitive domain can be visualised for social cognitive development. Figure 3.20 shows the relation between the cognitive domains of an individual cognitive agent and possible social cognitive domains. As can be seen, an emotion can become transferred among the members of a society as an extended emotion. When most people have that emotion via the extended emotion, it becomes an emotion held by the society. Similarly, when most people believe a particular cognitive pattern, it becomes the reality template of the society.

When people discover a new pattern via the abstraction of thought that cannot be accommodated within the existing socially accepted cognitive patterns, there is a paradigm shift, which results in changes in the socially accepted cognitive practices. Most people learn the new pattern via psycholinguistic metaphorical abstraction, as described in Section 3.5. An example is how Newtonian physics was replaced by relativity. Newtonian physics was a cognitive practice across society. Much of engineering and other disciplines were founded on Newtonian physics, which became a cognitive practice or reality template across society. However, Einstein identifies discrepancies in

Newtonian physics and proposes the theory of relativity. This was an example of abstract thinking by Einstein. Subsequently, many other people came to understand the new pattern and it became a social cognitive pattern.

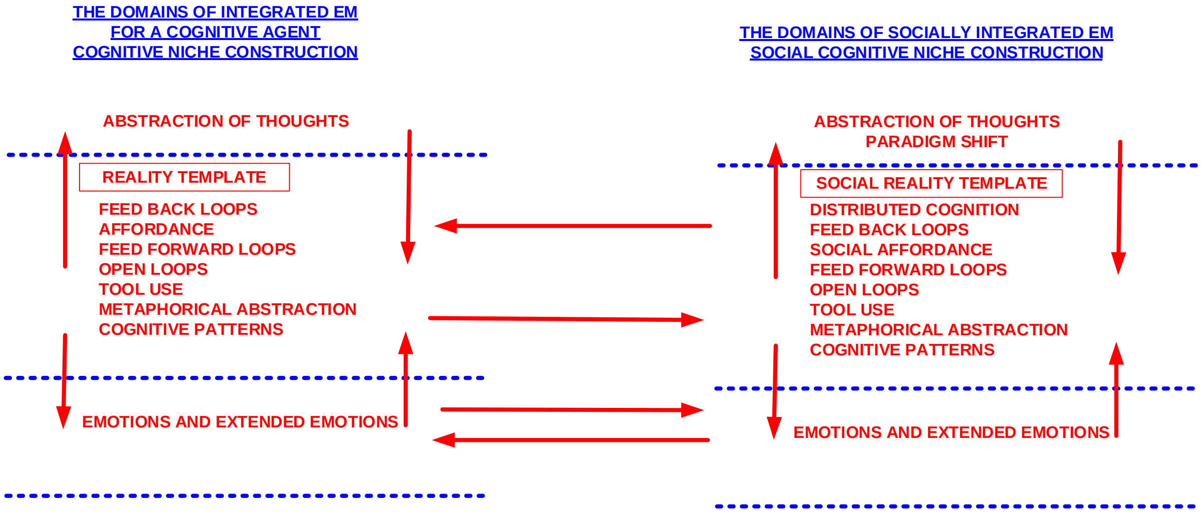


Figure 3.20. Relation between the cognitive domains of an agent and the domains in social cognitive development.

Figure 3.21 summarises the above discussion regarding the dynamic development of an individual psyche and its interactions with society. It is a two-way interaction. The bottom of the figure shows the direction of emotions and the abstraction of thought. There is continuity in the development of the abstraction of thought, just like Sterelny (2003) described for cumulative epistemic engineering. It has a horizontal arrow facing in the direction of time to indicate the progress in the abstraction of thought. In contrast, emotions are momentary; therefore, the arrow is perpendicular to time. Above the arrows for emotions and the progress of the abstraction of thought, the left of the figure shows the developmental aspects of the cognitive agent and how each individual cognitive agent feeds into the social cognitive domains at a particular time, which is represented by the triangle on the left. When a paradigm shift occurs, the social cognitive domains change, which is represented by the triangle on the right and is emerging from the left triangle.

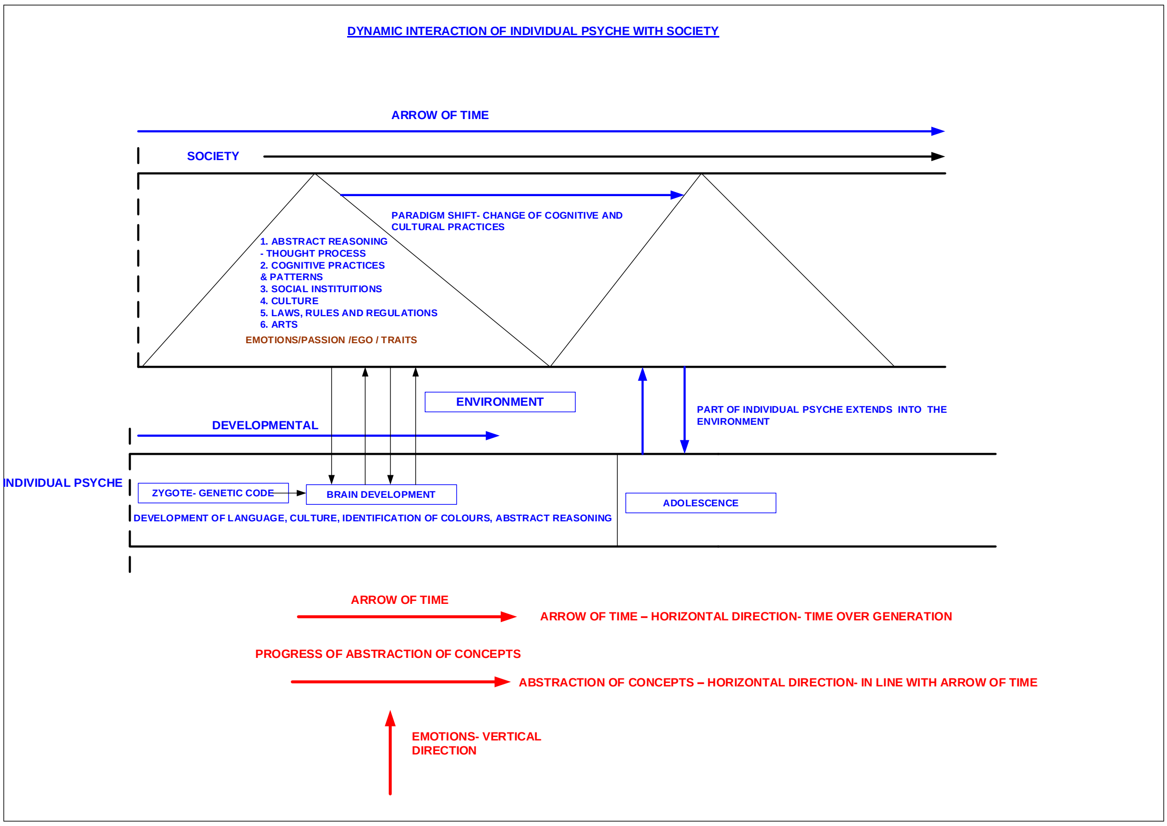

DYNAMIC INTERACTION OF INDIVIDUAL PSYCHE WITH SOCIETY
ARROW OF TIME
SOCIETY
PARADIGM SHIFT- CHANGE OF COGNITIVE AND CULTURAL PRACTICES
1. ABSTRACT REASONING
- THOUGHT PROCESS
2. COGNITIVE PRACTICES
& PATTERNS
3. SOCIAL INSTITUITIONS
4. CULTURE
5. LAWS, RULES AND REGULATIONS
6. ARTS
EMOTIONS/PASSION /EGO / TRAITS
ENVIRONMENT
PART OF INDIVIDUAL PSYCHE EXTENDS INTO THE ENVIRONMENT
DEVELOPMENTAL
INDIVIDUAL PSYCHE
ZYGOTE- GENETIC CODE
BRAIN DEVELOPMENT
ADOLESCENCE
DEVELOPMENT OF LANGUAGE, CULTURE, IDENTIFICATION OF COLOURS, ABSTRACT REASONING
ARROW OF TIME
ARROW OF TIME – HORIZONTAL DIRECTION- TIME OVER GENERATION
PROGRESS OF ABSTRACTION OF CONCEPTS
ABSTRACTION OF CONCEPTS – HORIZONTAL DIRECTION- IN LINE WITH ARROW OF TIME
EMOTIONS- VERTICAL DIRECTION

Figure 3.21. Dynamic development of an individual psyche and its interactions with society.

3.8.4 Conclusions

The modified Clarkian version of EM is immune to the coupling-constitution fallacy and cognitive bloat. It is clear that NCT not only substantiates the importance of feedback loops with the environment but also provides an explanation for the cognitive development of humans from an evolutionary perspective, such as cognitive niche construction. The application of dynamics to cognition not only supports feedback loops and the constitutive role of external artefacts but also explains the nature of cognitive agents over time. Developmental systems theory provides a systems approach to cognitive development and the nature of cognitive agents. Patterns in external information-bearing structures not only explain the nature and history associated with the engagement of an external artefact by a cognitive agent, but they also provide an explanation for the continuity of the development of artefacts and the progress in abstract reasoning embedded in the development of external artefacts. Such explanations offer historical insights into the cognitive development of humans, the nature of cognitive agents and the relations between cognitive agents and artefacts.

The nature and cognitive development of an agent and the relation of a cognitive agent to society were established by integrating a modified form of EM with NCT, CN and developmental systems theory.

This integrated version of EM and the broad patterns developed from the intersection of EM with NCT, DST and patterns from external information-bearing structures have the potential to support an integrated approach to multidisciplinary research.

4 ANTI-LUCK VIRTUE EPISTEMOLOGY

4.1 Introduction

This chapter describes a post-Gettier epistemology for analysing knowledge and explores a possible enrichment of post-Gettier epistemology based on the integrated version of an extended mind outlined in Chapter 3. A traditional analysis of knowledge, such as a justified true belief (JTB), has a tripartite structure. However, lucky true beliefs, as in Gettier-style cases, can undermine knowledge. Further, when an artefact is involved in cognitive processes, such as extended cognition, as discussed in Chapters 2 and 3, the risk associated with the artefact can undermine knowledge acquisition. The overall objective of the forthcoming chapters is to identify a suitable account of knowledge that can accommodate the extended cognitive processes detailed in Chapters 2 and 3.

The main objectives of this chapter are:

1. To identify an adequate theory of knowledge
2. To identify an account of luck that can adequately capture knowledge-undermining luck, especially when an artefact is involved in the cognitive processes.

This chapter mainly focuses on Gettier cases, various accounts of luck, anti-luck epistemology (ALE), robust virtue epistemology (RVE), anti-luck virtue epistemology (ALVE) and criticisms of ALVE.

The nature and characteristics of luck and its relation to knowledge are still being debated in epistemology. There are various theories of luck, such as probability, lack of control account, the modal account of luck (MAL) and various hybrid accounts. I am going to argue that none of them is an adequate theory of luck. However, based on Gettier cases, there is a consensus among epistemologists that knowledge is incompatible with at least some kinds of luck, such as veritic luck, which I am going to describe later in this chapter. Some forms of luck, such as a serendipitous event like the accidental discovery of penicillin, are compatible with knowledge. In this chapter, I am going to evaluate various accounts of luck, and thereafter, I am going to discuss the kinds of luck that are relevant in epistemology, i.e. the kinds of luck that undermine knowledge acquisition. Pritchard proposed ALE with a safety condition based on MAL for degettierising beliefs. In an analysis of knowledge, RVE considers that knowledge is an achievement due to the exercise of the cognitive ability of an epistemic agent. RVE has merits over other rival theories of knowledge. RVE, as a complete theory of knowledge, can address Gettier-style cases. However, there is a dilemma regarding whether RVE can capture the safety of the target proposition in all cases. Additionally, RVE may be too strong in cases of testimony. Pritchard argues that a safe true belief is not sufficient for knowledge. Pritchard accepts that knowledge is cognitive success due to ability. Therefore, Pritchard begins with the view that the ability and safety conditions are the two master intuitions that should guide our theorising about knowledge. The chapter then examines ALE, which has a safety condition based on MAL to address Gettier cases. Such cases occur when someone has a belief that is true by luck rather than through reliable cognitive abilities. I also discuss RVE, which views knowledge as an achievement from an individual's cognitive abilities. RVE is particularly compelling as it can address Gettier-style cases effectively, but it also faces challenges, such as its applicability to knowledge gained from testimony. The dilemma arises when considering whether RVE can account for the safety of knowledge in all situations. To address this, Pritchard combines the insights of ALE and RVE into a new framework called ALVE, which integrates the idea that knowledge is not just about cognitive success due to ability (a core concept of RVE) but must also include a safety condition so that it is immune to luck (as emphasised in ALE).

To ensure the safety of the target belief against knowledge-undermining luck, Pritchard (2016a) proposes ALVE with a condition for safety in addition to the ability condition. There are criticisms of ALVE by various people, such as Greco, Sosa, Kelp and Carter. All these philosophers agree with Pritchard that knowledge must be suitably non-lucky and that it must arise from ability. However, the core of the dispute concerns the idea that knowledge is success from ability. If this is satisfied, does that ensure that a belief is also non-lucky in the way that knowledge demands? After a careful evaluation of Pritchard's ALVE and various criticisms, I conclude that Pritchard's MAL is not an adequate account of luck and that ALVE is not an adequate theory of knowledge. The various alternative proposals by Greco, Sosa, Kelp and Carter also fail as an adequate theory of knowledge.

Section 4.2 focuses on traditional epistemology, Gettier cases, the philosophy of luck, ALE and issues with ALE. Section 4.3 details virtue epistemology (VE), ALVE and criticisms of ALVE.

4.2 Gettier Cases, Philosophy of Luck and ALE

The traditional analysis of knowledge is reductive, i.e. it relies on basic concepts without involving knowledge to avoid circularity. It provides individually necessary and jointly sufficient conditions for knowledge. In the traditional reductive analysis of knowledge, an agent *S* knows a proposition *p* is true if and only if:

1. *p* is true.

2. *S* believes that *p* is true.

3. *S*'s belief that *p* is true is justified.

Therefore, the traditional analysis of knowledge is based on such JTBs (Gettier, 1963).

Gettier (1963) provided some counterexamples to the traditional analysis of knowledge. In these cases, a subject's justified belief is, in fact, true but the JTB does not suffice for knowledge. Russell's stopped clock is an example of a Gettier case. Note that this example is not due to Gettier. By looking at a clock, John believes that the time is 7.30 am. John has good reason to believe that the time is 7.30 am and it is true that it is 7.30 am. However, unbeknownst to him, John was looking at a stopped clock. In this case, John's belief is true and justified but it cannot count as knowledge because John's belief is true due to luck. Such Gettier-style examples have a JTB, but the true belief is too lucky to count as knowledge. They are, thus, counterexamples to the classical account of knowledge. Such Gettier-style cases demonstrate that the traditional tripartite structure of knowledge is prone to knowledge-undermining luck.[2]

Here are some more examples of Gettier cases.

Roddy

> Roddy is a farmer. One day he is looking into a field near-by and clearly sees something that looks just like a sheep. Consequently, he forms a belief that there is a sheep in the field. Moreover, this belief is true, in that there is a sheep

[2] Note that Gettier never used the word "luck" in his paper. Gettier took it as intuitive that the counterexamples he gave were not cases of knowledge. Subsequent authors have attempted to explain why the cases he used as counterexamples do not produce knowledge. One common explanation is that, in these cases, it is only a matter of luck that the belief that each subject forms is true (Pritchard, 2015a); thus, the beliefs are unsafe.

> in the field in question. But what Roddy is looking at is not a sheep, however, but a big hairy dog that looks just like a sheep and which is obscuring from view the sheep standing just behind. (Pritchard, 2016a, p. 7)

This case is due to Chisholm (1977).

Temp

> Temp's job is to keep a record of the temperature in the room that he is in. He does this by consulting a thermometer on the wall. As it happens, this way of forming his beliefs about the temperature in the room will always result in a true belief. The reason for this, however, is not because the thermometer is working properly, since in fact it isn't – it is fluctuating randomly within a given range. Crucially, however, there is someone hidden in the room next to the thermostat who, unbeknownst to Temp, makes sure that every time Temp consults the thermometer the temperature in the room is adjusted so that it corresponds to the reading on the thermometer. (Pritchard, 2016a, p. 38)

Barney

> Barney is driving through the county and happens to look out of the window into a field. In doing so, he gets to have a good look at a barn-shaped object, whereupon he forms the belief that there is [a] barn in the field. This belief is true, since what he is looking at really is a barn. Unbeknownst to Barney, however, he is presently in "barn façade county" where every other object that looks like a barn is actually a convincing fake. Had Barney looked at one of the fake barns, then he would not have noticed the difference. Quite by chance, however, Barney just happened to look at the one real barn in the vicinity.
> (Pritchard, 2016a, p. 8).

This example is due to Carl Ginet (1975) and then most famously by Goldman (1976).

In these examples with Roddy, Temp and Barney, the agent forms a true belief, but it cannot be counted as knowledge because it was down to luck. The true belief so formed is fragile as the agent could have formed a false belief in a modally close possible world. Gettier cases show that, in addition to JTB, knowledge requires a further condition that blocks knowledge-undermining luck. Pritchard (2007, 2016a) proposes ALE to address Gettier cases. As per Pritchard, ALE has three parts: (1) a theory of luck, (2) the delineation of luck from knowledge to show that luck undermines knowledge and (3) an anti-luck condition.

Lucky true beliefs raise concerns about the attribution of knowledge. Therefore, it is necessary to characterise the nature of luck and its relation to knowledge in knowledge attribution. The objective in analysing knowledge is to define knowledge unambiguously. A traditional analysis of knowledge has a tripartite structure, such as in JTB. However, as we have seen, Gettier-style cases show that JTB is insufficient for knowledge. Gettier cases show that, in addition to JTB, knowledge needs to satisfy a further condition that blocks knowledge-undermining luck. Therefore, Gettier-style cases show that JTB is insufficient for knowledge. Note that not all types of luck are incompatible with knowledge. Pritchard (2005) classifies luck as:

1. Veritic luck: It is a matter of luck that the agent's belief is true. In veritic luck, an agent's belief is true in the actual world but false in nearby possible worlds. Veritic luck includes intervening luck, as in the case of Roddy, and environmental luck, as in the case of Barney. In intervening veritic luck, luck intervenes between an agent and the belief the agent forms. For example, a sheep-shaped hairy dog intervened to form Roddy's belief about sheep. Again, in intervening luck, the agent's belief is true in the actual world but false in nearby worlds. If there is environmental veritic luck, then due to the bad environment, the agent's true belief that was formed in the actual world could be wrong in nearby possible worlds. For example, Barney identified by luck a true barn in fake barn county. Barney could look at a fake barn in a modally close possible world and form a false belief about the barn. Because of luck, Barney saw a real barn in fake barn county and formed a true belief about the barn. Intervening luck, such as with Roddy, and environmental luck, such as with Barney, are incompatible with knowledge. Therefore, the luck that undermines knowledge is mainly intervening or environmental luck.
2. Reflective luck: Given only what the agent can know by reflection alone, it is a matter of luck that their belief is true.

3. Evidential luck: Serendipitous discoveries, such as the discovery of penicillin, are an example of evidential luck. This type of luck is compatible with knowledge, as the subject has properly gained JTB.[3]

As with the traditional structure of knowledge, true beliefs are a common factor for attributing knowledge and for lucky beliefs. However, for veritically lucky true beliefs, justification is not adequate for attributing knowledge. As per Pritchard, knowledge requires a true belief that is non-veritically lucky. Therefore, veritic luck is important in the attribution of knowledge. In these examples with Roddy and Barney, the agent forms a true belief that cannot be counted as knowledge because it was due to luck. Both intervening and environmental luck are outside agential capacity, as they do not depend on the agent and are objective. Therefore, it can be concluded that knowledge-undermining luck, i.e. veritic luck, is objective. If luck is anchored in the environment and independent of the agent's ability, such as with a sheep-shaped object in Roddy's case or the fake barns in Barney's case, it is objective. Subjective luck, on the other hand, relates to the agent's perspective or internal states. For example, if someone makes a correct guess based on an unfounded belief or hunch, that could be considered a case of subjective luck. The "luckiness" here is tied to the individual's internal perspective or reasoning process, not external factors. In the context of my argument, I focus on objective luck, especially as it relates to knowledge-undermining factors in Gettier cases.

I agree with Pritchard that veritic luck can undermine knowledge. Now the question is whether veritic luck covers the entire spectrum of luck that undermines knowledge, especially when an artefact is involved in knowledge acquisition? Pritchard's formulation of veritic luck does not specifically address luck when an artefact is involved in knowledge acquisition. When an artefact is involved in cognitive processes, especially extended cognition, the risk associated with the artefact is significant if the extended cognitive process results in knowledge. I think that such situations require an assessment of the various philosophies of luck and an evaluation of whether such accounts of luck can adequately address the entire spectrum of knowledge-undermining luck, especially when an artefact is involved in knowledge acquisition. Since knowledge can be

[3] Pritchard countenances further types of luck such as: Content epistemic luck: It is lucky that the proposition is true. Capacity epistemic luck: It is lucky that the agent is capable of knowledge. Doxastic epistemic luck: It is lucky that the agent believes the proposition. Evidential epistemic luck: It is lucky that the agent acquires the evidence that she has in favour of her belief. Note that these types of luck are not relevant for this thesis.

undermined by luck, luck has to be analysed to understand its relation to knowledge. In this section, I outline various accounts of luck, such as probability, MAL and a hybrid account, each of which can partially capture the nature of luck.

Two aspects of luck need to be evaluated:

1. What is the nature of luck? Is it best captured by probability, MAL or a hybrid account?
2. How can luck exclude knowledge?

My strategy here is first to evaluate the theories of luck and then engage with how knowledge can exclude luck by using the resources of the relevant account of luck.

In the following section, I discuss various philosophies of luck, such as the probabilistic account, MAL and various hybrid accounts. The latter combine components of the probabilistic and modal accounts. As discussed earlier, not all kinds of luck are incompatible with knowledge. For example, evidential luck is compatible with knowledge. In the next section, I will discuss the philosophy of luck and thereafter I will discuss the nature of luck that excludes knowledge.

4.2.1 Philosophy of Luck

The predominant views on the philosophy of luck are:

1. Probability account of luck
2. Lack of control account
3. MAL
4. Hybrid account of luck

I use two criteria to evaluate the various accounts of luck. One is whether luck is objective, and the second is whether luck can be measured. Since luck is beyond an agent's ability, I consider that luck is objective.

4.2.1.1 Probability Account of Luck

The probability account of luck considers a lucky event to be improbable. The luckiness of an event can be expressed in terms of probability. Luck is directly proportional to its probability.

As per Rescher (2014), luck has two extremes: a favourable outcome (lucky) and an unfavourable outcome (unlucky). The extent of luck is the difference between the span of luck and unluck:

> One key factor in determining luck is the difference between the actual outcome and what might have been: if the outcome is favorable, the agent is lucky to the extent that this result differs from its unfavorable alternative; if the outcome is unfavorable, the agent is unlucky to the extent that this differs from what would have been if things had gone well. Either way, the difference in value between an unfavorable and a favorable outcome is crucial for the extent of luck. The second key determinant of luck is probability. An agent is the more lucky not only with a favorable result that is of greater value but also with one that is more unlikely. (Rescher, 2014, p. 624)

Rescher (2014) claims that luck can be measured by the following equation. Let p be the probability of success and $1 - p$ the probability of failure. Δ is the difference in the values of the favourable and unfavourable outcomes. Then, the amount of luck λ can be expressed as follows:

Favourable result: $\lambda^{+} = \Delta \times (1 - p) = \Delta - \Delta p$

Unfavourable result: $\lambda^{-} = -\Delta p$

Rescher argues that if we consider the

> failure of probability (1 − p) as a measure of the risk, and the difference between a favorable and an unfavorable outcome (Δ) as a measure of the stake, then the amount of (good) luck at issue with a favorable result is simply the product of these two quantities risk × stake. (Rescher, 2014, p. 624)

As per Rescher (2014), in simple terms, with risk (R) as the probability of failure in a chancy situation and the stake (S) as the difference between a favourable and an unfavourable outcome, luck (L) can be calculated as the product of these quantities: $L = R \times S$.

There are counterexamples for the probability account of luck. For events with the same probability, luck can be intuitively different. Say the probability that a pair of night vision goggles fails is 0.001. This implies that out of 1000 instances of using these goggles, on average, they would fail once. This statistic is likely derived from testing or historical data and reflects the reliability of the goggles over many uses. In a lottery where the probability of winning is also 0.001, then if someone were to play the lottery 1000 times, she would expect to win once, on average. According to the probability account of luck, both these events – the goggles failing and winning the lottery – should be equally lucky (or unlucky), since they share the same probability. In a lottery, small changes (like choosing just one different number) can lead to drastically different outcomes, namely winning or losing. This contrasts with the goggles, as failure is not typically due to one minor factor but a culmination of issues or a significant defect. The probability of 0.001 in both scenarios indicates a rare event. For the goggles, the probability indicates their reliability over time and the aggregation of experiences. By contrast, in the lottery, the probability reflects the immediate chance of a rare event occurring in a single instance. This perception occurs partly because a very small change (such as choosing one or two different numbers) can be the difference between winning a substantial prize and winning nothing. The chance of winning is purely random, as it depends on the draw of the numbers. There is no accumulation of risk or reliability over time; each lottery draw is a separate event with the same odds. In summary, a probability of 0.001 for both the night vision goggles and lottery cases indicates a rare event, but the implications, interpretations and perceptions of this probability differ significantly due to the different natures of these events.

The probability of failure of night vision goggles (0.001) could be the same as the probability of buying a winning lottery ticket (0.001), but these events do not have intuitively similar luck. Winning a lottery can mean selecting just the few correct digits. In addition, the one winning lottery ticket among 1000 tickets is immediate. However, the immediate failure of any night vision goggles is remote. The following examples from De Grefte (2020) also show this. Jill's company wants to fire one of its 1000 employees. It has assigned a unique number to each and will pick a number at random, and that person will be fired. Therefore, Jill's probability of being fired is 1/1000 = 0.001. Consider another case. Joe has to present an important document at a meeting today. If he fails to attend the meeting with the document, he will be fired. To prevent this, Joe has put extensive safeguards in place. He has already put the document in his briefcase, added a reminder in Alexa and told his wife to remind him about the document and the meeting. Even with such safeguards, the probability that Joe will fail to attend the meeting with the document is 0.001. Ultimately, Jill and Joe both kept their jobs. Intuitively, Jill is luckier than Joe although the probability of failure was the same for both. For Jill, the probability of being fired is entirely dependent on an external random process (a lottery), over which she has no control. This randomness and lack of agency contribute to the perception that Jill, by not being selected for termination, is particularly lucky. Joe's situation, in contrast, involves active measures taken by him to avoid an unfavourable outcome. Regardless of the identical probability of failure, Joe's extensive safeguards (putting the document in his briefcase, setting reminders, etc.) imply a level of control and personal agency over the situation. When he successfully keeps his job, it is perceived as less a matter of luck and more a result of his preventative actions. The comparison to a lottery, in which winning is seen as highly lucky due to its randomness, further underlines that perceptions of luck are influenced by more than just probabilities. Thus, the probability account of luck cannot capture this intuition of luck.

Rescher's equation that quantifies luck is very similar to the quantification of risk in engineering, which is applicable for the failure of an external artefact such as night vision goggles. As per ISO/IEC standards, risk is defined as "the combination of the probability of occurrence of harm and the severity of that harm. The probability of occurrence includes the exposure to a hazardous situation, the occurrence of a hazardous event, and the possibility to avoid or limit the harm" (IEC 61511, 2016, p. 25).

The quantification of risk in engineering:

$$\text{Risk} = \text{Probability of a failure event occurring} \times \text{Severity of harm}$$

which can be simplified as:

$$\text{Risk} = P_f \times E(C)$$

where P_f is the probability of a failure event and $E(C)$ is the undesired consequence.

Comparing Rescher's equation with risk in engineering, R in Rescher's equation is the same as P_f, as it is the probability of failure. The stake S in Rescher's equation is the difference between a favourable outcome and an unfavourable outcome. It is a measure of the difference in value between outcomes. In contrast, $E(C)$ in engineering risk is specifically focused on the severity or extent of harm or negative consequence resulting from an event. This is a more targeted measurement and is typically concerned with harm, loss or damage. Although, there is a similarity in the structure of the two equations, they serve different purposes and contexts. Rescher's equation deals with the broader concept of luck, which can encompass both positive and negative outcomes and their impact on an individual’s situation. Engineering risk, on the other hand, is specifically concerned with the likelihood and impact of negative events (failures), often in a more quantifiable and tangible sense. Therefore, even though there are parallels in how probability and impact are considered in both equations, the scope and focus of Rescher’s equation and those of the engineering risk equation are distinct.

I am interested in comparing these equations because, in engineering, the quantification of failure based on probabilities is well established for engineering components, namely external artefacts such as night vision goggles. Therefore, in cognition involving such an external artefact, quantifying the risk for the knowledge gained when using the artefact is important. As an example, the failure of night vision goggles could cause false beliefs or lucky true beliefs, depending on the failure mechanism of the goggles.

The probability view of luck is like the risk assessments used in engineering. Interestingly, the counterexamples against the probability account of luck are applicable in risk assessments in engineering, as the complete spectrum of risk cannot be captured by probability. There are risk cases in engineering that cannot be captured merely by using probability. I will explore this in detail in Chapter 5 when I discuss the risks associated with artefacts.

4.2.1.2 Lack of Control Account of Luck

As the term indicates, the lack of control account of luck (LCAL) considers that luck is outside the control of the agent. For epistemic luck, i.e. knowledge-undermining luck, LCAL is very much aligned with the virtue reliabilist theory of knowledge in which knowledge is cognitive success due to the ability of the agent. Therefore, LCAL has two complementary ideas: (1) luck is not within the agent’s ability in the acquisition of knowledge and (2) luck is, generally, beyond the agent’s control. These are not exactly the same. The first concept particularly relates to the attribution of knowledge, as shown by counterexamples to virtue reliabilism. A notable example is environmental luck, such as in the barn façade scenario, in which an agent’s belief is true due to factors outside their control and ability. Another example of intervening luck is Archie’s success in shooting an arrow. One gust of wind blew it off target, but a second gust sent it back on course, showing that luck is outside his ability. The second idea is about the metaphysics of luck. Riggs argues that:

> One has control over some happening to the extent that the happening is properly considered something the agent has *done*. First, the event has to be the product of the agent's powers, abilities, or skills. Second, the event has to be, at least in some attenuated sense, something the agent *meant to do*. This second requirement does not demand an actual conscious intention on the part of the agent, but it does mean that a goal or desire or intention must be guiding the exercise of one's powers, abilities or skills that brings about the event in question. (Riggs, 2009, p. 11)[4]

In the initial formulation of LCAL by Riggs, there was no significance condition. In that formulation, luck was conceptualised as being due to events outside the control of an agent. This broad definition inadvertently included virtually any event beyond an individual's control, regardless of its impact on the person. Under this definition, everyday occurrences like sunrise, which are clearly outside any human control, could technically be categorised as instances of luck. This is because the event (sunrise) meets the sole criterion of being beyond the agent's control. Pritchard argued that according to LCAL, events like the sunrise, which are outside an individual's control, would be classified as instances of luck, which contradicts our intuitive understanding of what constitutes luck. Mundane events like sunrise, which is outside the control of an agent, should count as luck according to LCAL. Therefore, Riggs adds a significance condition to LCAL (Riggs, 2009, p. 17):

E is lucky for *S* if and only if:

(a) *E* is (too far) out of *S*'s control, and
(b) *S* did not successfully exploit *E* for some purpose, and
(c) *E* is significant to *S* (or would be significant, were *S* to be availed of the relevant facts).

Pritchard (2014) later drops the significance condition from MAL due to agent subjectivity and the potential for pragmatic encroachment in the ascription of luck. Later, Riggs (2014) admits that the addition of a significance condition to LCAL was hasty and

[4] For the relation between knowledge and luck, Riggs argues that: "I have defended a theory of knowledge according to which *S* knows that *p* so long as *S*'s believing the truth about *p* is not a matter of luck. Luck, in turn, was defined in terms of the extent to which an agent has control over an outcome. Hence, *S* knows that *p* if and only if *S*'s believing the truth about *p* is an outcome that is/was sufficiently under *S*'s control. This kind of view is sometimes referred to as a 'control theory' of luck" (Riggs, 2014, p. 628).

unnecessary. Riggs (2014) provides an example that shows that events out of an agent's control, such as a solar eclipse, can be lucky, if the situation necessitates. For example, Jones and Smith are explorers and are captured by a tribe in Africa. The people decide to execute them. Jones knows that a solar eclipse is due and that the tribal people will not execute them if there is an eclipse. However, Smith is completely unaware of the solar eclipse and the tribal customs. Both keep their life due to the solar eclipse. Riggs argues that for Smith, the solar eclipse is a lucky event but that it is not lucky for Jones. Pritchard criticises this, since luck is objective and not prone to a subject's attribution of luck. It should not matter what Smith or Jones thinks about the luckiness of a solar eclipse. Since a solar eclipse is outside an agent's control, as per LCAL, a solar eclipse can be counted as lucky, which is counter-intuitive.

Lackey develops counterexamples to show that LCAL is neither necessary nor sufficient for luck. Some mundane events are outside an agent's control but are significant to an agent. For example, if my spouse collects our child from school, that is significant to me, but it is outside my control and is not lucky. However, as per LCAL, this event is lucky, which is counter-intuitive. This is against the sufficiency condition of LCAL. To challenge the necessity condition, Lackey provides a counterexample of a demolition worker who successfully completes the wiring for the destruction of a building. Unbeknownst to him, a mouse chews through the wiring, which disconnects the circuit. Before the worker presses the switch, his colleague hangs a jacket on a nail, which reinstates the circuit. The demolition worker then presses the switch, and as expected, the building is demolished. Lackey argues that: "What demolition worker shows, then, is that although an event may be within a given agent's control, that the agent has such control can itself be largely a matter of luck, and hence the event resulting from this control can be lucky as well" (Lackey, 2008, p. 259). Lackey's counterexample is similar to Archie above, where the agent has sufficient control but this control is almost interrupted by factors unbeknownst to the agent. Finally, through a combination of purely coincidental and unlikely factors, the agent's control is not, in fact, interrupted, so that the agent's control is riddled with luck, which, in turn, extends to the resulting event. This is a counterexample to LCAL (Lackey, 2008). Lackey, thus, argues that LCAL is not a necessary or sufficient condition for luck.

Riggs (2014) argues that luck is a mere coincidence in the acquisition of knowledge as a matter of degree, rather than all or nothing. According to Rigg: "Luck, in turn, was defined in terms of the extent to which an agent has control over an outcome. Hence, *S* knows that *p* if and only if *S*'s believing the truth about *p* is an outcome that is/was sufficiently under *S*'s control" (Riggs, 2014, p. 628). However, it is not clear what "control" means nor its relation to the outcome. For example, in Chapter 3, I discuss the interaction of a cognitive agent with the environment and the establishment of various loops, such as open loops, feedback loops, feedforward loop, etc. The agent has control of these various loops during their interaction with the environment based on parameters such as vision, hearing, engagement with various artefacts, etc. However, it is not clear in LCAL how abilities, control and the extent of control define luck.

I take the liberty of defining LCAL based on the control loops described in Chapter 3. Consider the case of Barney. His barn-spotting ability is based on an open loop and has fixed parameters, i.e. Barney's barn-spotting ability, which is based on an open-loop interaction, produces true beliefs in a normal barn environment. In fake barn county, Barney's barn-spotting ability is still based on an open loop with fixed parameters. However, Barney can form a false belief about a barn by looking a fake barn. Here, luck is beyond Barney's control. But it is not clear from LCAL how luck associated with the environment interacts with the agent's ability and control. However, such events cannot be considered as lucky. Lackey's demolition counterexample shows that an event can be lucky even if the agent has control over the outcome. Based on the above, I conclude that LCAL is not an adequate account of luck.

4.2.1.3 Modal Account of Luck

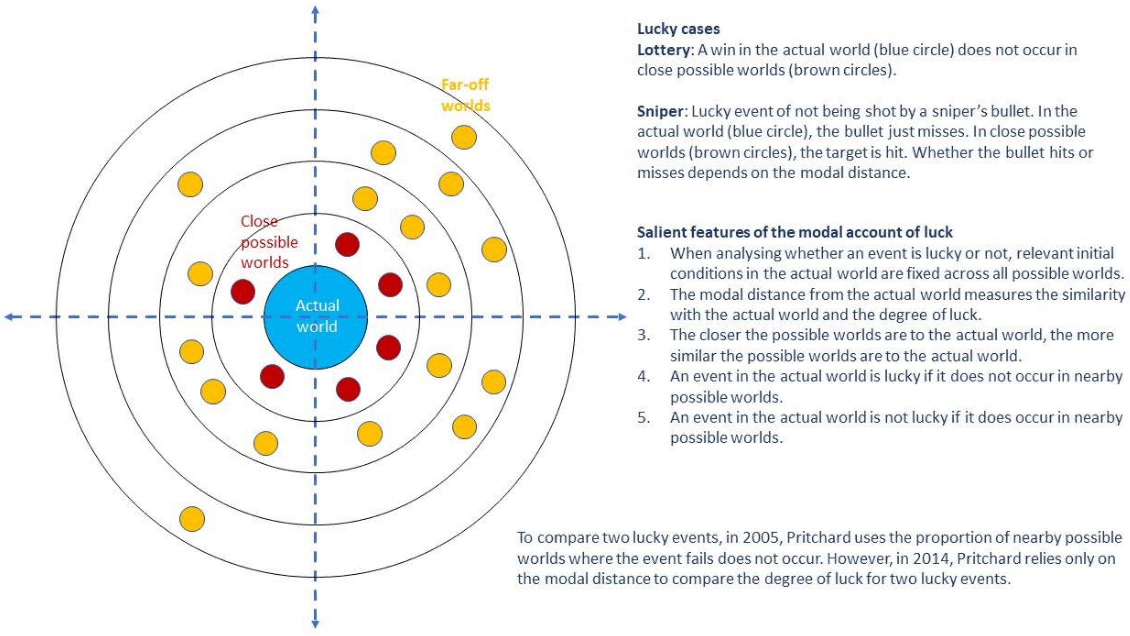
Pritchard (2005) describes MAL as follows:

> (L1) If an event is lucky, then it is an event that occurs in the actual world but which does not occur in a wide class of the nearest possible worlds where the relevant initial conditions for that event are the same as in the actual world.
> (L2) If an event is lucky, then it is an event that is significant to the agent concerned (or would be significant, were the agent to be availed of the relevant facts). Though vague, this condition should suffice to capture the basic contours of the "subjective" element of luck, and thus also capture the sense in which luck can be either good or bad.
>
> (Pritchard, 2005, pp. 125, 132)

As per MAL, a lucky event is an event that occurs in the actual world but does not occur in a wide class of nearby possible worlds. In terms of epistemic luck, a lucky true belief is true in the actual world but false in some close possible worlds, if the relevant initial conditions for the event are fixed to those in the actual world. For example, sunrise is not a lucky event, as it happens in the actual world as well as in a wide class of nearby possible worlds. However, winning a lottery in the actual world is a lucky event, as it does not happen in a wide class of nearby possible worlds. Pritchard noted two important complexities associated with MAL (Pritchard, 2007):

> (1) The class of relevant possible worlds must be restricted to those where the initial conditions for the target event are the same as those in the actual world. For example, it does not make sense to consider a nearby close world with a losing lottery ticket if the agent does not buy a lottery ticket in that world.
>
> (2) Luck will come in degrees depending on how close the actual world is to nearby possible worlds where the event does not occur. This indicates that some lucky events are luckier than others.

Figure 4.1 illustrates MAL. The blue circle at the centre is the actual world. The outer amber and brown circles are possible worlds.



Lucky cases
Lottery: A win in the actual world (blue circle) does not occur in close possible worlds (brown circles).

Sniper: Lucky event of not being shot by a sniper's bullet. In the actual world (blue circle), the bullet just misses. In close possible worlds (brown circles), the target is hit. Whether the bullet hits or misses depends on the modal distance.

Salient features of the modal account of luck

1. When analysing whether an event is lucky or not, relevant initial conditions in the actual world are fixed across all possible worlds.
2. The modal distance from the actual world measures the similarity with the actual world and the degree of luck.
3. The closer the possible worlds are to the actual world, the more similar the possible worlds are to the actual world.
4. An event in the actual world is lucky if it does not occur in nearby possible worlds.
5. An event in the actual world is not lucky if it does occur in nearby possible worlds.

To compare two lucky events, in 2005, Pritchard uses the proportion of nearby possible worlds where the event fails does not occur. However, in 2014, Pritchard relies only on the modal distance to compare the degree of luck for two lucky events.

Figure 4.1. Modal account of luck.

To evaluate whether an event is lucky or not, the relevant initial conditions in the possible nearby worlds have to be fixed to those in the actual world. The modal distance from the actual world indicates the extent of the similarity with the actual world and the degree of luck. The closer the possible worlds are to the actual world, the more similar they are to the actual world. An event in the actual world is lucky if the event does not happen in nearby close possible worlds. An event in the actual world is not lucky if it happens in most of the possible nearby worlds. To compare two lucky events, Pritchard (2005) uses the percentage of nearby possible worlds where the event fails to happen. However, since the attribution of luck is objective, Pritchard (2014) drops significance condition L2. Thus, in 2014, Pritchard uses only the modal distance to compare the degree of luck for two events. The modal account of epistemic luck focuses on a target event – "the forming of a belief about whether *p*" – for which the potential for the knowledge acquired in the actual world can be undermined in a nearby possible world by veritic epistemic luck.

Pritchard (2014) argues that it is clear that there is not a linear relation between the probability of a lucky event and the modal distance of the event. Probability refers to how likely an event is to occur, whereas modal distance refers to how "close" an event is to happening in the nearest possible worlds. Pritchard suggests that our intuition of luckiness depends on how modally close the lucky event is. He compares playing the lottery with winning gold in the 100 m in the Olympics. Both are probabilistically unlikely. They may be remote for the vast majority of us. Although both events are lucky, winning the lottery is modally close while winning gold in the Olympics is remote. Winning the lottery is highly improbable (low probability) but modally close. In the nearest possible worlds, very little would need to change for someone to win the lottery, perhaps just one different number being drawn. Winning an Olympic gold medal in the 100 m is also improbable for most people, but it is also modally distant. In the nearest possible worlds, significant changes would be required for the average person to become an Olympic gold medallist (years of training, physical attributes, etc.). Pritchard infers that since our intuition about luck is influenced more by modal closeness than mere probability, similarly, our judgement of knowledge (or the reliability of a belief) could also be more influenced by how resistant that knowledge is to error across possible worlds, rather than just the statistical probability of being correct in the actual world. From this, Pritchard rightly concludes that our judgement of a piece of knowledge is sensitive to the modal closeness of error as opposed to the probabilistic closeness. Pritchard argues that there is an epistemic continuum of modally close worlds with respect to a target event. Luck can be perceived in degrees, such that the luckiness of an event varies continuously from 0% to 100%. Near to 0% indicates that luckiness is low in the possible world and near to 100%, that luckiness is high.

The objective of the modal assessment is to ensure that the knowledge (cognitive success) acquired by an agent is a true belief and unlucky. Thus, MAL considers the modal profile of a target event such that cognitive failure occurs due to veritic epistemic luck if the relevant initial conditions are fixed. As explained above, there is an epistemic continuum for the luckiness of a target event. If the target event, i.e. cognitive failure, occurs in a modally close world, then the cognitive success is due to veritic epistemic luck and cannot count as knowledge. If the cognitive failure is modally far off, the cognitive success can be counted as knowledge.

MAL, thus, uses a continuum of epistemic luck in which one end is modally close to luck, which is incompatible with knowledge, and the other end is modally distant from epistemic luck, which is compatible with knowledge. To assess safety, both the initial conditions for the lucky event and the belief-forming process must be constant (basis-relative account of safety). Consider Barney. The fake barns with a single true barn and Barney's perceptual apparatus for belief formation are fixed in both the actual world and each modal world. In the actual world, by luck, Barney sees a real barn and forms a true belief about the barn. In a modally close world, Barney could have looked at a fake barn and formed a false belief about the barn.

In 2014, Pritchard argues that:

> The degree of luck involved varies in line with the modal closeness of the world in which the target event doesn't obtain (but where the initial conditions for that event are kept fixed). We would thus have a *continuum* picture of the luckiness of an event, from very lucky to not (or hardly) lucky at all.
>
> (Pritchard, 2014, p. 600)

De Grefte (2020) suggests a counterexample to Pritchard's MAL to show that the degree of luck does not depend exclusively on the modal distance:

> *Case 1*: Jaimy buys the single winning lottery ticket out of 10,000,000 tickets.
>
> *Case 2*: Jerome buys the single winning lottery ticket out of 100 tickets.

In both cases, the lottery ticket is selected in the same way. According to MAL, both Jaimy and Jerome are lucky, since in a nearby close world, they would have lost if just a single digit on the ticket were different. Thus, the modal distance for Jaimy and Jerome is the same and so, per MAL, the degree of luck is the same for them. However, intuitively, Jaimy is luckier than Jerome. MAL cannot capture this difference; therefore, the modal distance is not useful for quantifying luck. De Grefte's (2020) counterexample shows that modal distance alone is not adequate for differentiating the extent of luck in the two lotteries.

As Pritchard (2007) rightly concludes, there are complexities with MAL. Important aspects are: (1) How should the relevant initial conditions be fixed across the possible worlds? (2) The degree of luck depends on how close the nearest possible world is in which the target event does not occur (Pritchard, 2007, p. 31).

4.2.1.4 De Grefte's Hybrid Account

In the above example of Jaimy and Jerome, the modal distance cannot account for the degree of luck; however, a probabilistic theory of luck can differentiate between the luckiness of Jaimy and Jerome. Thus, Jaimy can be considered luckier than Jerome, as the probability of Jaimy winning was 1/10,000,000 but the probability for Jerome was 1/100. However, for Jill and Joe, probability theory cannot differentiate between their luckiness as the probabilities of them losing their jobs are the same, although intuitively, Jill is luckier than Joe. Therefore, the probability theory of luck alone cannot differentiate their luckiness. For Jill and Joe, the modal distance can explain the degree of luckiness. Although Jill's situation is undoubtedly a matter of luck due to its dependence on a random process, the modal distance involved is greater than in scenarios in which fewer variables can alter the outcome. The multitude of possible combinations in the random draw suggests that there are many alternative scenarios in which Jill retains her job, implying a significant modal distance. For Jill to lose her job, a specific sequence of digits must align. Jill's scenario involves multiple variables (digits), making the specific event of her losing her job less modally close. However, for Joe, several things have to change in nearby possible worlds; therefore, being sacked is modally remote compared with the situation for Jill.

Since neither modal distance nor probability alone can be used to compare two lucky events, De Grefte (2020) proposes a hybrid account of luck that has modal and probabilistic components:

> *Luck 1*: The degree of luck (partially) depends on probability. Other things being equal, the lower the probability of an event, the higher the degree of luck.

> *Luck 2*: The degree of luck (partially) depends on modal distance. Other things being equal, the closer the nearest world where the event fails to occur is to the actual world, the higher the degree of luck.

De Grefte considers that it is best not to give a special weighting for either probability or modal distance, so both are given an equal weight. It is clear that neither of the theories can completely characterise luck. Each of them has an intuitive appeal in partially addressing luck but both fail to provide a complete theory of luck. However, MAL is superior to the probabilistic account in characterising luck qualitatively. MAL is adequate for evaluating a single event qualitatively using the modal distance, since there is a decrease in similarity with the actual world as the modal distance increases and a corresponding increase in luckiness as the modal distance decreases. However, there are limitations in comparing the luckiness of two events, as in the example of Jaimy and Jerome. Moreover, fixing the relevant initial conditions across the possible worlds has a direct impact on the modal assessment of luck. Variations in fixing the initial relevant conditions can result in changing a lucky event to a non-lucky event. MAL requires a way to compare the luckiness of two events and another to fix the initial relevant conditions across possible worlds.

4.2.2 Epistemic Luck

After evaluating various theories of luck, I conclude that none of them can adequately explain luck. MAL has merits over other accounts of luck and can evaluate the luckiness of a single event. However, De Grefte's counterexample effectively challenges MAL's ability to differentiate degrees of luck between different but similar single events. In the example, Jaimy buying a winning lottery ticket from a pool of 10,000,000 tickets and Jerome buying a winning ticket from a pool of 100 tickets are both deemed equally lucky by MAL due to having the same modal distance. However, intuitively, Jaimy seems luckier. This example shows that MAL, although useful in assessing whether an event is lucky, struggles to quantify the degree of luck in various scenarios. In summary, as mentioned earlier, MAL is adequate for establishing the luckiness of a single event qualitatively. Since the relation between knowledge and luck relies on evaluating the luckiness of a single event, problems in comparing two lucky events will not undermine the application of the modal intuition to epistemic luck. In terms of epistemic luck, as I explained earlier, luck attribution is objective in line with knowledge attribution. If that is the case, then it is necessary to analyse in detail belief-forming processes to fix the relevant initial conditions across the nearby possible worlds. This analysis helps in establishing the relevant initial conditions in the nearest possible worlds, which is a key factor in assessing the modal closeness and, consequently, the luckiness of a belief. By understanding the nature of the belief-forming process (whether it involves open loops, feedback loops, etc.), we can better assess how likely it is for a belief to remain true or become false in close possible worlds. Thereby we can evaluate its susceptibility to epistemic luck.

MAL is superior to other accounts when qualitatively evaluating the epistemic risk. However, as we have seen earlier, to capture the complete spectrum of risk, especially when an artefact is involved in knowledge acquisition, an account of luck requires both probabilistic and modal components. Such a hybrid account of luck can capture the full spectrum of risk. A hybrid account of luck requires a probability assessment in addition to the modal distance to compare the luckiness of two events. Epistemic luck involves an element of chance in acquiring true knowledge. When external artefacts are involved, their failure probability introduces an aspect of luck into the cognitive process. For example, if a scientific instrument has a high probability of error, any knowledge derived from it is subject to a higher degree of luck, namely the chance that the instrument might not provide accurate data. On the other hand, a low probability of error suggests a more reliable process with less luck.

A hybrid account of luck integrates both a probability assessment and the modal distance. The modal distance indicates how easily an event could have turned out differently, whereas the probability aspect quantifies the likelihood of an event occurring. This dual approach allows for a more comprehensive assessment of luck, as it accounts for both the likelihood of an event and its stability across possible worlds. The probability aspect of the hybrid account can be used to assess the risk or likelihood of failure of artefacts. For example, the probability of a technological malfunction or error in an instrument directly impacts the reliability of the cognitive process and the knowledge derived from it. From an epistemological perspective, this approach underlines the importance of considering both the likelihoods of success and failure (probability) and the stability of outcomes (modal distance) in assessing the validity of knowledge obtained through external means.

In conclusion, neither MAL nor the probability account of luck are an adequate theory of luck. An adequate theory of luck requires both the probabilistic component and the modal component.

4.2.3 Anti-Luck Epistemology

As Riggs (2014) points out, there is a general consensus among epistemologists that luck undermines knowledge. This section provides the motivation for ALE (Pritchard, 2016a). The Gettier-style cases demonstrate that the traditional tripartite structure of knowledge is prone to knowledge-undermining luck. Based on the philosophy of luck, Pritchard proposes ALE to address Gettier-style cases in which luck undermines knowledge. Pritchard's ALE is based on MAL. As per Pritchard, ALE has three parts: (1) a theory of luck, (2) the delineation of luck from knowledge to show that luck undermines knowledge and (3) an anti-luck condition. Here Pritchard evaluates whether safety or sensitivity is better for a modal assessment. The attempts to define modal conditions for luck, such as sensitivity and safety, have succeeded in addressing some of the Gettier cases. According to the sensitivity principle: "One has a true belief such that, had what one believed been false, one wouldn't have believed it" (Pritchard, 2015a, p. 99). Sensitivity captures the modal condition for luck. Recall the stopped clock case where John forms a true belief that the time is 7.30 am. John has good reason to believe that the time is 7.30 am, and it is true that it is 7.30 am. However, unbeknownst to him, John was looking at a stopped clock. In this case, John's belief is true and justified but it cannot count as knowledge because John's belief is true due to luck. Now, apply the modal condition of sensitivity. In the closest possible world, the clock would still be stopped, the time would be slightly different, and there would be nothing to indicate that the clock is stopped. In this closest possible world, John also believes that the time is 7.30 am but the belief is false. Therefore, John's belief is insensitive. Modal conditions based on sensitivity can address Gettier cases. However, sensitivity fails to capture inductive knowledge, as with Ernie.

Ernie

Ernie lives in a high-rise block of flats in which the way to dispose of one's garbage is to drop it down a garbage chute in the corridor. Ernie knows that the flats are well maintained, and so when he drops his garbage down the chute, he believes that it will soon be in the basement (Pritchard, 2012). Since the building is well maintained, normally the garbage bag will end up in the basement. However, when applying the sensitivity condition, Ernie's belief is insensitive, such that, in close possible worlds, Ernie's belief is false, as the garbage can become stuck although Ernie continues to believe that the garbage bag is in the basement. Although Ernie has not seen the bag arrive at the basement, his belief is based on induction.

The anti-luck condition for safety has merits over the sensitivity condition as it can capture inductive knowledge. Although Ernie's belief is insensitive, it is safe, since in all close possible worlds, as the building is well maintained, there is no reason for the bag to become stuck and it will reach the basement. Ernie will continue to believe that the bag is in the basement and his belief will be true. The discussion here focuses on a modal ALE based on safety. Luck will come in degrees depending on the proximity of the nearest possible world in which the target event does not occur. The closer the nearest possible world in which the target event does not occur, the luckier the event will be. Safety is the preferred condition for ALE. Safety demands a true belief, not just in the actual world but also in close possible worlds. Pritchard proposes that a useful way of thinking about this is in terms of risk. Applied to the epistemic case, the risk of a false belief in the lottery case is much higher, since it is modally closer than for Ernie, and this has a bearing on our willingness to attribute knowledge. According to modest ALE, although the anti-luck condition is necessary for knowledge, it will not be sufficient for knowledge (with a true belief).

Robust ALE, in effect, treats the anti-luck intuition as the dominant intuition when analysing knowledge in terms of a true belief that satisfies the relevant anti-luck condition, i.e. the safety principle. However, consider the Temp case we discussed earlier, in which knowledge is intuitively absent but the belief is safe. If the hidden helper continues to have the same role in the closest possible world as in the actual world, then Temp's belief is guaranteed to be true in the closest possible world. Obviously, Temp cannot get knowledge by consulting a broken thermometer; however, his belief is true and safe in the actual world and in the closest possible world. Pritchard (2016a) argues that:

> The direction of fit between belief and fact in this case is all wrong. What we want in a case of knowledge is for one's beliefs to be responsive to the facts. In this case, however, the direction of fit is entirely in reverse, since the facts are in effect responding to the agent's beliefs rather than *vice versa*.
>
> (Pritchard, 2016a, p. 39)

Temp's belief about the temperature is true and safe (it remains true in the closest possible worlds) due to the intervention of a hidden helper who adjusts the broken thermometer. This creates a situation in which the belief is immune to epistemic luck, as it is guaranteed to be true in the closest possible worlds. Although Temp's belief is true and safe, the cognitive success (the accurate belief about the temperature) is not attributable to Temp's cognitive agency. Instead, it is the result of external manipulation by the hidden helper. The direction of fit between belief and fact is reversed; the facts (the temperature reading) are responding to Temp's belief (or situation), rather than Temp's belief being responsive to the actual temperature. If Temp has full control over the belief-formation process, then the belief could indeed qualify as knowledge. However, this immunity to luck is externally imposed and not a result of Temp's cognitive control or agency. This external imposition undermines the belief's qualification as knowledge, despite it being immune to epistemic luck. If Temp's belief was immune to luck due to his own cognitive control – for instance, if he had developed a reliable method for verifying the temperature – then the belief could be considered knowledge under ALE.

The Temp case poses a problem for ALE. Robust ALE fails if a belief is radically immune to knowledge-undermining epistemic luck as this true belief does not qualify as knowledge because this cognitive success is in no way significantly attributable to the subject's cognitive agency, such as with Temp.

Pritchard proposes ALE to ensure the safety of a true belief against knowledge-undermining luck. However, Pritchard realises that, in some cases like Temp, true beliefs are safe but that epistemic success cannot be attributed to the agent. Therefore, ALE is not an adequate theory of luck. In short, there are cases of belief that satisfy the anti-luck intuition but which do not satisfy knowledge attribution. Therefore, Pritchard concludes that, as a theory of knowledge, ALE has limitations. Note that the overall objectives of this chapter are to (1) to identify an adequate theory of knowledge that can accommodate cognitive processes involving external artefacts and (2) to identify an adequate account of luck in relation to knowledge when knowledge acquisition involves artefacts.

In the following section, I discuss a theory of knowledge, VE, that considers that knowledge is a cognitive success due to the exercise of the cognitive ability of the agent.

4.3 VE and ALVE

VE considers that knowledge is a cognitive success due to the exercise of the cognitive ability of the agent. According to Greco, knowledge is success from ability or competence exercised by the agent's own cognitive agency. Greco (2009) defines knowledge as follows:

> *S* knows *p* if and only if *S* believes the truth (with respect to *p*) because *S*'s belief that *p* is produced by intellectual ability. The term "because" is here intended to mark a causal explanation. The idea is that, in cases of knowledge, the fact that *S* has a true belief is explained by the fact that *S* believes from ability. (Greco, 2009, p. 18)

Pritchard (2010b) argues that:

> How are we to read the "because of" relation here? There is as yet no consensus amongst robust virtue epistemologists on this score, but the most developed view in the literature in this regard due to Greco (2007) takes the causal explanatory line that true belief is because of an agent's cognitive abilities when it is primarily creditable to the agent that her belief is true. (Pritchard, 2010, p. 26)

Note that the idea that knowledge must arise from ability is weaker than the claim made by RVE proponents that knowledge is an achievement, which implies that the cognitive success is primarily creditable to the exercise of cognitive ability as opposed to merely significantly creditable, as is maintained by modest forms of VE. Knowledge acquisition by an agent is a success due to an ability, as acquiring knowledge is a kind of achievement. Greco (2009) argues that RVE has considerable explanatory power and can address standard Gettier cases. Greco (2003) claims that strange and fleeting processes are not integrated with cognitive agency, such that any success due to them cannot be considered as an achievement due to ability. The cognitive success (acquiring a true belief) should emerge from a belief-forming process that is deeply rooted in the agent's cognitive character. This cognitive character includes their intellectual virtues: stable, reliable faculties or abilities for forming true beliefs. It is the belief-forming process, grounded in the agent's cognitive character, that needs to be integrated with the cognitive success for it to be considered a genuine cognitive achievement.

If knowledge is success from ability, then according to Greco, *S* has a true belief formed by competent cognition. In Gettier cases, *S* has a true belief. *S*'s belief is formed by competent cognition, but *S* does not have knowledge because her belief is not *primarily* due to ability. *S*'s forming a true belief is merely lucky (Greco, 2020, p. 89). This is very clear from Roddy, who has excellent vision. Roddy is looking into a field and clearly sees something that looks just like a sheep. Consequently, he forms a belief that there is a sheep in the field. It is a true belief, as there is a sheep in the field. Note that Roddy exercises competent perception and ends up with a true belief, but this true belief is not because Roddy has exercised competent perception. Instead, it is down to luck that there is a sheep in the field, unseen and unknown to Roddy. The true belief cannot be attributed as a success due to his ability. Thus, Roddy formed the true belief that there is a sheep in the field by mere luck.

RVE (Pritchard, 2016a) insists that only those reliable belief-forming processes that make up one's cognitive character, such as one's cognitive faculties, cognitive abilities and intellectual virtues, can generate knowledge. A cognitive success cannot be considered as knowledge if it is not related to the agent's cognitive abilities, even if the belief is reliably produced and justified. Pritchard argues that, as mentioned above, Temp's belief about the temperature in a room cannot be considered as knowledge although it was reliably produced. Temp lacks knowledge because his belief about the temperature is not a product of his cognitive ability, as it was guaranteed by the hidden helper. Temp cannot gain knowledge by looking at a broken thermometer. For Temp, as Pritchard (2012) argues, the direction of fit between belief and fact is wrong. In knowledge acquisition, the agent's belief is responsive to the facts, but for Temp, the facts, i.e. the temperature of the room, respond to Temp's beliefs. Temp's true belief is not sufficiently creditable to his cognitive agency but instead, is more due to some feature of the situation that is completely unconnected with his cognitive agency.

Pritchard agrees that RVE, e.g. as per Greco, can get the right result in standard Gettier cases with intervening luck, given that in such cases the target belief is not correct because of ability, such as Roddy, whose beliefs about sheep in the field were reliably formed by his cognitive abilities but with the help of some epistemic luck, since he was looking at a sheep-shaped object rather than a real sheep. In this case, Roddy lacks knowledge about the target proposition although his belief was formed due to his reliable abilities. As discussed earlier, Roddy exercises competent perception and forms a true belief but does not do so because he exercised competent perception. Instead, it is down to luck that there is a sheep in the field, unseen and unknown to Roddy. The true belief cannot be attributed as a success due to his ability. Instead, Roddy formed the true belief that there is a sheep in the field by mere luck. Thus, his epistemic success cannot be attributed to his cognitive abilities. RVE can address this issue by giving credit primarily to cognitive abilities. The merit of this approach is that it does not require a non-virtue theoretic condition to ensure safety. Therefore, it can be concluded that virtue reliabilism can address standard Gettier cases with intervening luck.

VE has merits over other rival theories of knowledge. VE, as a complete theory of knowledge, can address Gettier-style cases. As we have seen earlier, simple safety-based theories of knowledge can also address Gettier-style cases. However, such theories cannot ensure cognitive success through the ability of the agent. In contrast, cognitive success in VE is due to the ability of the agent. However, there is a dilemma regarding whether VE can capture the safety of the target proposition in all cases, such as for Barney. In the closest possible world, Barney looks at a fake barn and forms a false belief about it. Barney's cognitive ability cannot ensure the safety of the target proposition. Additionally, VE may be too strong in cases of testimony where the hearer can gain knowledge but exercise a minimal level of cognitive ability and the credit for the hearer's knowledge is primarily attributed to the exercise of cognitive abilities by the speaker. Lackey (2009) raises a dilemma relating to RVE based on testimonial knowledge, which is knowledge that has been acquired by a hearer that is mainly attributed to the speaker's ability rather than the hearer's.

Jenny

Jenny gets off a train in an unfamiliar city and asks the first person that she meets for directions. The person that she asks is indeed knowledgeable about the area and helpfully gives her directions. Jenny believes what she is told and goes on her way to her intended destination (Pritchard, 2012, p. 269).

There is, thus, a dilemma for VE. A strong reading of VE, i.e. by giving credit primarily to cognitive abilities, can address standard Gettier cases, like Roddy; however, to accommodate testimonial knowledge, such as Jenny, a weak reading of VE is required, i.e. significant credit is given to cognitive abilities rather than credit being given primarily to cognitive abilities. Although Roddy exercises his cognitive ability, his true belief cannot be attributed as knowledge because it was not primarily due to his cognitive ability. In contrast, Jenny exercises a minimal degree of cognitive ability to acquire knowledge and her cognitive success is not primarily due to her exercise of cognitive ability. This raises a dilemma. The strong reading of VE, i.e. that cognitive success is primarily due to the cognitive abilities of the agent, can address Gettier cases; however, for testimonial knowledge, the cognitive success is not primarily due to the cognitive ability of the hearer. The strong reading of VE does not allow the attribution of knowledge due to testimony, but the weak reading of VE fails to address Gettier cases with intervening luck.

Pritchard (2012) argues that RVE is too strong in testimonial situations where the information given to the hearer cannot be attributed as knowledge, as the hearer gains knowledge that is not primarily creditable to their cognitive ability. The utterings cannot be considered as knowledge because the information is fully dependent on the speaker's cognitive character, such as when seeking directions from a stranger in an unfamiliar city. Pritchard further argues that RVE faces a problem with environmental luck, such as the example of barn façades. Barney truly believes that he is looking at a real barn after employing his cognitive abilities, for example, by looking directly at a barn. Since Barney's true belief is solely due to his cognitive ability, his cognitive success cannot be called knowledge given that his belief could so easily have been false. This necessitates some modal conditions (safety) to avoid epistemic luck. From the barn façade case, Pritchard concludes that knowledge and cognitive achievements are not the same thing, as Barney makes a cognitive achievement but his belief cannot be considered knowledge. Therefore, RVE is not a viable theory of knowledge.

In summary, Pritchard (2007) argues that a safe true belief is not sufficient for knowledge. Pritchard agrees that knowledge requires that one's cognitive success is due to ability, but cognitive success due to ability alone cannot ensure the safety of the beliefs formed. Therefore, ALE and RVE cannot be considered as adequate theories of knowledge. Pritchard (2012) claims that neither ALE nor VE suffices to capture ability and safety, thus the need for ALVE. To ensure the safety of the target belief against knowledge-undermining luck, Pritchard (2012) proposes ALVE with a condition for safety in addition to the ability condition.

4.3.1 Anti-Luck Virtue Epistemology

There are cases of belief that satisfy the anti-luck intuition but which do not satisfy the ability intuition. Therefore, Pritchard (2007) concludes that, as a theory of knowledge, ALE has limitations in accommodating the ability intuition. In the previous section, we note that Pritchard concludes that RVE is based on the cognitive ability of the agent; however, it is prone to environmental luck, as in the barn façade case where cognitive success and achievement by the exercise of cognitive ability are compatible with environmental luck. Barney has cognitive achievement without knowledge. Therefore, RVE based on the ability intuition fails as a complete theory of knowledge. Knowledge is incompatible with luck. From the example of Barney, Pritchard (2012) argues that to have a complete theory of knowledge, a safety condition is required in addition to the ability condition. Pritchard argues that RVE with the ability intuition as a central thesis cannot accommodate the anti-luck intuition. Pritchard concludes that a theory of knowledge should include two master intuitions: (a) an ability condition to produce and justify knowledge and (b) a degettierised (anti-luck) condition to overcome knowledge-undermining luck such as environmental luck. Pritchard argues that each condition should have the same weight, in the sense that they each provide a fundamental intuition about knowledge. Pritchard believes that such a new theory of knowledge that can accommodate both the ability condition and the anti-luck condition can answer Gettier-style cases as well as the value problem of knowledge without considering that knowledge is, finally, valuable. Pritchard, therefore, proposes a complete theory of knowledge, ALVE, which requires: (1) an ability intuition (VE) and (2) an anti-luck safety condition, as in ALE.

As per Pritchard (2012), the general structure of ALVE is as follows:

> Knowledge is *safe belief* that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is to a *significant degree creditable* to one's cognitive character. The safety element of the view is the anti-luck condition, while the virtue-theoretic clause is the ability condition. (Pritchard, 2012)

Notice that ALVE does not demand that the cognitive success in question must be *because of* the agent's cognitive ability, only that it should be to a significant degree creditable to her cognitive character, which is a weaker claim.

In Pritchard's ALVE, Jenny's true belief is partly creditable to her cognitive abilities (but not primarily creditable) and it is safe knowledge. Therefore, Jenny's belief can be considered as knowledge. Roddy's and Barney's beliefs cannot be considered as knowledge. Although their beliefs are produced by reliable cognitive abilities and traits, they do not meet the anti-luck condition of ALVE, i.e. the agent's true belief could very easily have been false. As discussed earlier for Temp, his belief is safe and it is guaranteed to be true because of the hidden helper:

> While this belief satisfies the anti-luck intuition, it does not satisfy the ability intuition. In short, while Temp's beliefs aren't luckily true – in fact, they are *guaranteed* to be true – this doesn't in any substantive way reflect his

> cognitive agency, but is rather attributable to the external intervention of the hidden helper. (Pritchard, 2016a, p. 39)

Pritchard claims that ability and safety are the two master intuitions required for knowledge. Accordingly, Pritchard proposes ALVE, which captures both the agent's ability and safety in knowledge acquisition.

4.3.2 Criticisms of ALVE

The requirement for these two independent master intuitions, i.e. the ability intuition and an anti-luck safety condition, has been challenged by various people, such as Greco, Sosa, Kelp and Carter. All these philosophers agree with Pritchard that knowledge must be suitably non-lucky and that it must arise from ability. In this respect, these philosophers agree with Pritchard's master intuitions. However, there is a dispute about the idea that knowledge is success from ability. If this is satisfied, does that ensure that a belief is also non-lucky in the way that knowledge demands?

4.3.2.1 Greco

Greco (2007), for example, thinks that if the agent satisfies the ability condition for knowledge, then this will entail that the agent's belief is also not going to be lucky in a way that is incompatible with knowledge. Here, Pritchard disagrees with Greco. Greco argues that VE can provide better insights into the relation between luck and knowledge. Knowledge is success from ability so is different from mere lucky success.

Greco (2010) states that the abilities needed to achieve results have certain properties and a structure:

(1) They are reliably successful in some way relevant to the ability in question.
(2) Abilities are tied to relevant conditions.
(3) Abilities are always relative to the environment.
(4) Abilities have a good track record with respect to achieving a result (Greco, 2010, p. 61).

Summarising the above, Greco (2010) argues that "*S* has an ability *A*(*R*/*C*) relative to environment *E* = Across the set of relevantly close worlds *W* where S is in *C* and in *E*, *S* has a high rate of success in achieving *R*" (Greco, 2010, p. 61). Now, Greco realises that such an account of knowledge-based ability faces a generality problem:

> Depending on how we specify the relevant *W*, *C*, and *E*, we will get variable success rates with respect to *S*'s believing the truth. For example, it may be that *S*'s success rate is very high if conditions are specified more narrowly, but very low if conditions are specified more broadly. (Greco, 2010, p. 61)

To avoid a generality problem with ability, Greco relies upon the concept of knowledge proposed by Edward Craig, which flags good information and good sources of information for use in practical reasoning. Greco (2010) argues that combining the properties and structure of ability with this concept of knowledge will provide enough resources to clarify the relation between luck and knowledge where "relevant parameters of ability should be specified according to the interests and purposes of relevant practical reasoning" (Greco, 2010, p. 61).

Greco (2012) responds to the barn façade case by considering ability relative to the normal environment that serves the practical purposes. In the barn façade case, Barney's barn-spotting ability is not in a normal environment. Accordingly, Barney's barn-spotting ability cannot be considered as a cognitive success. Greco argues that: "*S* has a knowledge-relevant ability *A* (*R/C/D*) relative to an environment *E* = *S* has a disposition to believe truths in range *R* when in circumstances *C* and environment *E*, with degree of reliability *D*" (Greco, 2012, p. 18). To avoid the generality problem with such a structure, i.e. the possibility of getting potentially different results from ability based on a narrow or wide specification of the relevant environment (*E*), circumstances (*C*) and range of truths (*R*), Greco proposes that such parameters are set by the needs for information and information sharing for relevant practical tasks.

Barney (*S*) has an ability to spot a barn (*R*) relative to the environment *E* and conditions *C*. Based on the above, Greco concludes that Barney's barn-spotting ability is reliable in normal circumstances but not in the barn façade environment. Therefore, Barney does not know that the object he has seen is a barn. Greco's response, which emphasises the reliability of cognitive abilities in normal circumstances, does not fully address the issue of environmental luck in scenarios like the barn façade case. Although Greco acknowledges the limitations of cognitive abilities in such environments, this does not completely resolve Pritchard's concern about the role of luck in undermining knowledge. Pritchard argues that, particularly in the barn façade case, even when a cognitive ability is exercised and leads to a true belief (cognitive success), this does not necessarily equate to knowledge. Barney's true belief about seeing a barn is more a result of luck than his cognitive abilities, due to the deceptive environment with fake barns.

Greco tries to address the challenges raised by Jenny's case. In VE, the exercise of an agent's cognitive ability is necessary for cognitive success or knowledge acquisition. However, in testimonial cases like Jenny, the credit for the cognitive success is not given to Jenny but to the reliable informant. The standard objection is that VE is overly individualistic and cannot accommodate social dimensions of knowledge, such as social epistemic dependence in knowledge acquisition such as in testimony. Thus, Greco proposes two ways to address Jenny's case.

Greco considers a non-epistemic analogy, namely a soccer game in which Ted, a soccer player, receives a brilliant pass and scores a goal. Here Greco compares it with the cognitive success of Jenny. Greco (2010) argues that if Ted is passive and does not receive the pass, he would not score a goal. In this case, Greco argues that Ted was involved in the right sort of way so as to get credit. Greco claims that:

> Credit for success, gained in cooperation with others, is not swamped by the able performance of others. It's not even swamped by the outstanding performance of others. So long as one's own efforts and abilities are appropriately involved, one deserves credit for the success in question. (Greco, 2010, p. 65)

Greco also used a strategy with an attribution relation as pragmatic:

> The purposes of practical reasoning are well served by the reliable reception of testimony and expert testimony. That is, in cases of testimonial knowledge, *S* has the right sort of ability, and employs it in the right sort of way, so as to serve the purposes of practical reasoning, i.e., those of *S* and those of the group that needs to depend on *S* as a source of good information. (Greco, 2010)

I agree with Pritchard:

> No one is denying that Jenny's cognitive success is in part due to her cognitive abilities: this case is not meant to be a counterexample to the ability intuition. The point, however, is that the degree of trust involved in this case means that the cognitive success is not *because of* Jenny's cognitive abilities – *namely*, it is not primarily creditable to Jenny's cognitive abilities that she is cognitively

> successful. To ensure that Jenny's cognitive success would be because of her cognitive abilities it would be necessary to substantially reduce this degree of trust, but then one is back to the original counter-intuitive response of arguing that Jenny lacks knowledge in this case after all. (Pritchard, 2010, p. 270)

Greco criticises ALVE, noting that knowledge requires ability that should be suitably non-lucky, that acquiring knowledge by ability ensures that epistemic success is non-lucky and that there is no requirement for an additional codicil in ALVE to ensure safety. It is clear that luck can be beyond an agent's ability, as for Barney, and that safety cannot be ensured by the way epistemic success is achieved by the cognitive ability of the agent. Therefore, the alternative proposals by Greco fail to provide an adequate theory of knowledge.

4.3.2.2 Sosa

Sosa (2007) also thinks that knowledge has to come from ability and must be suitably non-lucky. But Sosa's conception of what non-lucky involves (in a way that matters for knowledge) is compatible with a belief being unsafe; this follows from his idea about animal knowledge or an unreflective apt belief. Sosa differentiates knowledge into two varieties, namely animal knowledge and reflective knowledge. Reflective knowledge, as conceived by Sosa, involves a higher level of cognitive engagement. It not only requires the belief to be true and apt (arising from a reliable cognitive ability) but also necessitates that the agent has reflective awareness or understanding of the reliability and epistemic status of the belief. It is a more advanced form of knowledge, such that the agent critically evaluates and endorses their belief-forming processes. Sosa states that "animal knowledge is apt belief, which hits the mark of truth through the exercise of competence, intellectual virtues. Animal knowledge is essentially apt belief, as distinguished from the more demanding reflective knowledge" (Sosa, 2007, Ch. 2, p. 2). Sosa claims that ordinary perceptual beliefs thus retain their status as apt, animal knowledge (Sosa, 2007, Ch. 2, p. 2). Animal knowledge is an apt belief that is unreflectively justified, as with the perceptual belief of Barney. However, reflective knowledge, i.e. an apt belief aptly noted, is reflectively justified.

In Sosa's view, Barney has animal knowledge but he lacks reflective knowledge. Sosa's explanation may be counter-intuitive, as animal knowledge is possible in many Gettier examples where we cannot attribute knowledge. Therefore, Sosa's response does not adequately address Gettier cases like Barney.

4.3.2.3 Carter

Carter (2013a) also agrees with the two master intuitions but argues that there is a way of formulating an ability-based view such that satisfying the ability condition will ensure that a belief is suitably non-lucky. I critically evaluate two aspects here:

(1) Carter's argument against Pritchard's ALVE (Carter, 2013a)
(2) Carter's argument for RVE as ALE (Carter, 2016)

Carter (2013a) argues that the key premise of Pritchard's ALVE is based on the distinct nature of the ability condition and the anti-luck condition, such that neither of them can entail the other. This makes cognitive achievements compatible with knowledge-undermining environmental luck cases, such as Barney. Carter terms this the "independent thesis". Carter suggests that Pritchard's argument for the compatibility of a cognitive achievement with knowledge-undermining luck is incorrect. Carter attempts to prove that the independence thesis is wrong by establishing a belief-focused cognitive ability as a cognitive achievement and then proving that a cognitive achievement is not compatible with knowledge-undermining luck. The strategy Carter uses can be summarised as:

(1) Separate veritic luck from situational luck.
(2) Establish Pritchard's compatibility thesis in terms of a cognitive achievement, i.e. cognitive achievements are compatible with knowledge-undermining luck (veritic luck).
(3) Establish the difference between agent-focused success and belief-focused success. This involves distinguishing between success that is primarily attributed to the agent's abilities (agent-focused success) and success that centres on the reliability of the belief itself (belief-focused success). Agent-focused success emphasises the role of the agent's cognitive faculties in arriving at a true belief, whereas belief-focused success is more concerned with the truth of the belief, regardless of the agent's specific role in its formation.
(4) Distinguish cognitive achievements due to an agent-focused success (CA-As) from those due to a belief-focused success (CA-Bs). CA-A refers to achievements in which the agent's cognitive abilities play a central role in the formation of true beliefs. On the other hand, CA-B refers to situations in which the truth of the belief is paramount and the agent's specific cognitive contribution may be less central.

(5) Argue that Pritchard's aim was CA-A rather than CA-B but that VE theorists endorse CA-B. Carter's argument is that CA-Bs are not compatible with knowledge-undermining luck. He aims to show that a true belief (a cognitive achievement) formed through belief-focused success inherently counters the influence of veritic luck, thus aligning more closely with knowledge.

Thus, Carter (2013a) focuses on the distinction between CA-A and CA-B.

Carter argues that Pritchard considers CA-As as being at the core of VE. Carter further suggests that a CA-B, i.e. a belief-focused cognitive achievement, is at the core of VE since a CA-B matters for knowledge and is incompatible with knowledge-undermining luck.

Carter (2013a) argues that the core notion of VE theorists aligns with CA-B rather than CA-A, whereas to formulate ALVE, Pritchard interprets the core of VE as CA-A. Carter asks the following question: Is environmental luck, as in the case of Barney, really compatible with a cognitive achievement (CA-B)? Pritchard's focus was to characterise Barney's case as CA-A, such that the cognitive success of "believing *p* truly" is primarily creditable to Barney's cognitive ability. Carter proposes a negative, weaker condition for the exercise of cognitive ability in which a cognitive achievement is not primarily creditable to an agent's cognitive ability. Here is Carter's proposal:

> *Not primarily creditable (NC)*: That *S*'s *f*-aimed effort was successful is not primarily creditable to *S*'s *f*-abilities if, holding fixed the total contribution of *S*'s *f*-abilities manifested in the actual world in *S*'s attempt to bring about *f*, most nearby worlds are worlds where *S* fails to bring about *f*.

As per Pritchard, the achievements due to Barney's barn-spotting abilities are primarily creditable to Barney. This is CA-A, as it is due to an agent-focused belief. However, if we consider belief-focused success, then it is CA-B. Since there are fake barns, Barney's achievement in spotting a barn is not primarily creditable to Barney's barn-spotting ability.

Carter (2013a) argues that the above NC condition aligns well with the ordinary notion of success. Carter concludes that, by holding fixed the total contribution of Barney's cognitive abilities with respect to his truth-aimed barn belief, then most nearby worlds are worlds where the barn belief he forms is false. Barney's perceptual belief-forming skills are insufficient to ensure that the "target" of his belief is a real barn, rather than any of the nearby façades.

From Carter's proposed negative weak ability condition, which is a sufficiency claim for cognitive achievement, and from the example of Barney using the "not primarily creditable to ability intuition", then Barney attains cognitive achievement (NC) irrespective of environmental luck. CA-B is incompatible with luck since the cognitive success is not primarily creditable to the cognitive ability of the agent (NC). Therefore, Barney's cognitive achievement is incompatible with knowledge-undermining luck. In this case, Pritchard's argument for two distinct and mutually exclusive master intuitions of ability and anti-luck dissolves. Therefore, RVE can still be considered as a potential theory of knowledge without suffering from the problems raised by Pritchard. ALVE is not a viable theory of knowledge. Therefore, Barney's true lucky belief that there is a barn is not primarily creditable to Barney's cognitive abilities. Carter argues that: "The most plausible way to interpret the cognitive achievement thesis, then, is one according to which cognitive achievements are not compatible with environmental luck" (2013a, p. 274). This undercuts Pritchard's motivation for the leg of his independence thesis based on the ability and anti-luck conditions. Whether VE can accommodate the ability and anti-luck constraints in the analysis of knowledge is a live question.

The core issue here is whether VE can capture the safety of the target proposition. The characterisation of cognitive achievement as CA-A or CA-B does not ensure safety, as in the example of Barney. Knowledge cannot be attributed to Barney, as his true belief about the barn in a fake barn county is too fragile because, in the closest possible world, Barney looks at a fake barn and forms false beliefs about it.

Although Pritchard formulates ALVE based on two distinct master intuitions, namely safety (from ALE) and ability (from VE), the functioning of ALVE is not based on the independent thesis, as Carter claimed. ALVE is based on a safety/ability view where safety and ability are integrated.

The general perception of RVE is that a cognitive achievement based on beliefs formed by the exercise of an agent's cognitive ability is incompatible with knowledge-undermining luck. As with Barney, Pritchard identifies cases where the safety of the cognitive achievement of the agent due to the exercise of their cognitive abilities is not always preserved. Therefore, he concludes that cognitive achievements are compatible with knowledge-undermining luck. The criticisms levelled against Pritchard mainly attempt to reinstate the general perception of RVE that cognitive achievements based on a belief formed by the exercise of an agent's cognitive ability are incompatible with knowledge-undermining luck. Further, an inherent safety condition is available in the exercise of cognitive ability to avoid knowledge-undermining luck. As explained earlier, Pritchard argues that in RVE, achievement from an ability condition should be weak enough to capture Jenny's testimonial knowledge.

Carter (2016) argues that Pritchard and Kelp are looking for the material adequacy of the virtue condition. Carter states that the "wrong kind of fact problem" challenges RVE on a priori grounds. There is a logical gap between the kind of facts needed for virtue ability versus the kind of facts needed for safety. Carter proposes a new solution in which knowledge depends on various degrees of both ability and luck. The amount of ability required for knowledge gradually falls as the amount of luck increases. Thus, our cognitive success depends on ability, not instead of, but more so, than luck.

Carter (2016) further suggests that as a complete theory of knowledge, RVE can ensure the safety of beliefs formed that can count as knowledge through the ability condition. He indicates that there is a gap in the current form of RVE regarding whether the necessary condition of ability is sufficient for knowledge. Under what conditions can the correctness of *S*'s belief be said to depend sufficiently on *S*'s cognitive ability? Thus, the "wrong kind of fact problem" challenges RVE in defending entailment. If RVE is a complete theory, it should ensure the safety of the target belief formed by the exercise of ability. Carter terms this requirement the entailment principle:

> *Entailment*: If the correctness of *S*'s belief depends (sufficiently) on *S*'s cognitive ability, then *p*'s safety is ensured.

Carter claims that RVE must defend entailment, since it does ensure the safety of a cognitive achievement that is considered as knowledge. Carter summarises Greco's (2008, 2010, 2012a) proposals for the relation between ability and luck as a contrariety principle to address entailment, i.e. the relation between luck and ability in terms of success. If luck is the salient contributor to the success of a belief compared to ability, then the safety of the belief is undermined. If ability plays a salient role in success, the role of luck is nullified. Contrariety can be summarised as: "The attribution of a success to ability is *incompatible* with the attribution of that success to luck" (Carter, 2016, p. 145).

Carter argues that the contrariety principle is mistaken because it views ability and luck as mutually exclusive. For the golfer, which is an example of perception, Carter claims that there is a gradient relation between luck and ability. Carter argues that: “The right question then is not Greco’s – viz., whether my getting it right depends on luck or ability, but rather, where this dependence stands in the balance.” Carter devises a new solution to entailment based on the gradient between luck and ability:

> *Balance principle*: If the correctness of *S*’s belief that *p* depends (sufficiently) on *S*’s cognitive ability, then it depends on *S*’s ability more so than luck that *S*’s belief that *p* is true.

To check whether the balance principle can address Barney’s case, Carter argues that Barney’s belief that “there is a barn” in fake barn county is not due to Barney’s ability more so than luck. Here, luck and ability are in balance. If we hold fixed the perceptual abilities of Barney, then the true belief that Barney formed by looking at a real barn in fake barn county is not due to Barney’s ability more so than luck.

Thus, as above, knowledge depends on various degrees of both ability and luck. The amount of ability required for knowledge falls gradually as the amount of luck increases. Our cognitive success depends on ability, not instead of, but more so, than luck. Carter claims that his solution is better than Pritchard’s or Kelp’s, as there is no requirement for a codicil to RVE to ensure the safety of a belief.

Carter and Peterson (2017) give a counterexample to MAL using events *E* and *E** relating to a train journey from Edinburgh to London:

> Imagine that you are in Edinburgh but wish to meet up with your sister in London tonight. You decide to take the East Coast Express from Edinburgh to London. To your surprise, the train actually arrives on time at 7.59 pm. This event, call it *E*, is a lucky event because the East Coast Express usually arrives in London at least ten minutes after schedule. However, *E* is fairly close to being a non-lucky event. If the on-time performance had just been a little bit better, *E* would have occurred in “too many” nearby possible worlds and would thus not have counted as lucky. Now consider event *E**. This is the event in which you arrive in London no later than 7.59 pm. There is a very reliable express coach running from Peterborough to London that is scheduled to reach its destination before 7.59 pm. In a large number of the possible

> worlds in which you sit on the East Coast Express you notice that the train is running late as it stops in Peterborough. In those worlds you therefore transfer to the express coach and arrive on time in London no later than 7.59 pm.
>
> Is event *E** a lucky event? Our intuition is that it is not. *E** occurs in the same worlds as *E* and in every possible world located just a tiny bit further way. Those somewhat more distant worlds make the scales tip over. *E** is not a lucky event. … However, because *E* was in fact lucky, proponents of the modal account must concede that *E** is (by the same rationale) lucky too. But this is absurd. (Carter & Peterson, 2017, p. 2177)

Carter and Peterson (2017) conclude that: "The counterexamples outlined here are equally applicable if one accepts, as Pritchard (2014), Sosa (2015) and Carter (2016) have, that luck comes in degrees" (Carter & Peterson, 2017, p. 2178). Carter denounces the position that luck comes in degrees. I agree with Carter and Peterson that MAL cannot capture the full spectrum of luck that affects knowledge acquisition. The continuum of epistemic luck, i.e. the degree of epistemic luck that excludes knowledge, is not adequate for capturing luck. Therefore, Carter's assumption of the balance principle for knowledge and luck fails. It is clear that luck can be beyond an agent's ability, as for Barney, and that safety cannot be ensured by the way epistemic success is achieved by the cognitive ability of the agent. Therefore, the alternative proposals by Carter fail to provide an adequate theory of knowledge.

4.3.2.4 Kelp

Kelp (2013c) agrees with Pritchard, both on the point of the two master intuitions and also that satisfying one does not entail satisfying the other. Kelp differs with Pritchard about how to characterise the ability condition.

Pritchard (2010a) claims that "Temp's cognitive success is in no way a product of his cognitive abilities" (Pritchard, 2010a, p. 49). Unlike Pritchard, Kelp (2013) argues that the contribution of Temp's cognitive competences to his cognitive success parallels the contribution of Jenny's competences to her success almost exactly (Kelp, 2013, p. 270). Kelp (2012) notes that Temp has exercised his cognitive abilities by consulting the thermometer to acquire a belief about the temperature. Moreover, he may be sensitive to the temperature readings if they are significantly different from what his thermoreceptors are telling him about the room temperature. Kelp claims either (1) that if the contribution of Jenny's competences is substantive enough to render her success to a significant degree creditable to competence, then so is the contribution of Temp's competences or (2) that if the contribution of Temp's competences is not substantive enough to render his success to a significant degree creditable to competence, then neither is the contribution of Jenny's competences. Although Pritchard can handle either case separately, he cannot get both cases right simultaneously (Kelp, 2013, p. 270). Therefore, Pritchard's ALVE fails to capture the equivalence of Temp's case and Jenny's case.

To resolve this issue, Kelp combines Sosa's version of VE with Pritchard's ALVE to propose the safe-apt view. According to this view, knowledge is a safe and apt belief. Like Pritchard's ALVE, it has an ability condition (Sosa's VE) and a safety condition. Kelp claims that the safe-apt view can reap the benefits of Sosa's version of VE and Pritchard's version of ALVE while avoiding their drawbacks.

To devise a concept of knowledge that has both safety and a competence condition, Kelp (2013) follows a different strategy from Pritchard's ALVE. As noted above, Pritchard suggests that Craig's genealogy of knowledge supports the bipartite structure of ALVE. Kelp argues against the bipartite structure based on Craig's concept of knowledge and proposes an alternative bipartite structure, the safe-apt view of knowledge, which can accommodate ability and safety.

Kelp (2013) argues that Pritchard's bipartite structure depends on the distinction between an informant with reliable cognitive ability and an informant we can trust. Kelp (2013) proposes an improved version of VE, the safe-apt view, based on Sosa's manifestation of ability plus a safety condition. This alternative proposal is as follows. An agent is entitled not to inquire further into a question if we do not know the answer to it. Kelp asks what sorts of conditions would be required here and whether they are applicable when we do not know the answer. Kelp argues that the concept of knowledge should determine when a given agent is entitled to inquire no further into a question. Kelp concludes that a concept of knowledge with the function of marking when an agent is entitled to inquire no further into a question will feature both the safety and ability conditions. Thus, he establishes an independent reason for why the concept of knowledge should feature both conditions of the safe-apt view. According to the safe-apt view, knowledge has a bipartite structure with separate ability and safety conditions.

As per Pritchard, RVE fails to ensure safety for Barney. While the agent has cognitive achievement, this cannot be considered as knowledge due to environmental luck. Jenny acquired knowledge via testimony, but the cognitive achievement was primarily attributed to the speaker rather than the hearer. Therefore, a cognitive achievement is neither a necessary nor a sufficient condition for knowledge and RVE cannot ensure safety against veritic luck, such as environmental luck. Pritchard and Kelp argue that RVE is inadequate for ensuring safety. They assert that RVE is an inadequate ALE because a dilemma prevents RVE from adequately ensuring the safety of beliefs that satisfy virtue conditions.

The exercise of cognitive abilities and cognitive success in testimonial cases like Jenny is different from the exercise of cognitive abilities by Temp and the cognitive success of Temp in a manipulated environment. I disagree with Kelp that Jenny's and Temp's competences and cognitive successes are parallel and comparable. I agree with Kelp that Temp exercises his cognitive abilities, but his cognitive success in a manipulated environment (with a hidden helper) is not because of his exercise of cognitive abilities and it is not comparable to testimonial cases like Jenny. Jenny actively employs her cognitive abilities by choosing whom to ask for directions, assessing the reliability and trustworthiness of the information provided, and integrating this testimony into her existing knowledge to form a belief about the route.

In the original Temp case, as presented by Duncan Pritchard, Temp forms a true belief about the temperature by consulting a faulty thermometer, which, unbeknownst to him, is corrected by a hidden helper. Temp's belief is true and safe (not formed by luck), but Pritchard argues that the success is not due to Temp's cognitive abilities, as Temp's cognitive involvement is limited to reading the thermometer. There is no indication that Temp has any special skills or scepticism regarding the thermometer's accuracy. In a revised version of this case, Temp possesses the cognitive ability to read and even repair a faulty thermometer. He is also sensitive to temperature discrepancies. Despite his cognitive abilities, Temp's true belief about the temperature is still due to the external manipulation by the hidden helper, who corrects the broken thermometer. The revised case shows that even with cognitive abilities relevant to the task at hand, Temp's cognitive success (forming a true belief) is not due to these abilities. Instead, it is the result of external factors (the hidden helper). The revised Temp case challenges Kelp's safe-apt view because the safety of a belief is not intertwined with its aptness. This scenario suggests that safety and aptness, as conceptualised in the safe-apt view, may not always be as closely linked as the theory proposes. This emphasises the need for a more nuanced approach to accommodate the varying degrees of cognitive engagement and the influence of external elements in belief formation.

As explained earlier, RVE can address Gettier cases with intervening luck, such as the case of Roddy. However, RVE cannot address Barney's case adequately. See the previous discussion about Barney from Greco's and Sosa's perspectives. Greco responds to the barn façade case by considering ability relative to a normal environment. In the barn façade case, Barney's barn-spotting ability does not occur in a normal environment. Accordingly, Barney's barn-spotting ability cannot be considered as a cognitive success. This raises issues with the generality of specifying conditions and the environment too narrowly. In Sosa's (2009) response, Barney has animal knowledge but he lacks reflective knowledge. Sosa's explanation may be counter-intuitive, as animal knowledge is possible in many Gettier examples where we cannot attribute knowledge. There are problems with Greco's and Sosa's responses, and both lack a principled way to resolve the issue. From these discussions, I conclude that the barn façade case is a genuine problem for RVE, as environmental luck can undermine cognitive achievement.

Carter and Peterson's (2017) own counterexample against MAL shows the limitations of the balance principle when luck and knowledge are in balance. Accordingly, their assertion that Barney's true belief about the barn is not due to Barney's ability more so than luck also fails. In summary, none of the criticisms have been successful in establishing that RVE is adequate, that ability and safety should be retained, and that there is no requirement for ALVE. Clearly, environmental luck, as in the barn façade case, is an issue for ensuring safety in RVE. Therefore, an additional safety condition is required, as in ALVE. There is no additional merit in Kelp's safe-apt view compared with Pritchard's ALVE.

ALVE is based on MAL. However, there are counterexamples against MAL. The counterexamples raised by Carter and Peterson (2017) and by De Grefte (2020) show the limitations of MAL. Since Pritchard's MAL has limitations, then ALVE based on MAL is not adequate for capturing knowledge-excluding luck. Since MAL cannot capture the full spectrum of luck, ALVE cannot be considered to be an adequate theory of knowledge. I conclude that to capture the full spectrum of luck, both probabilistic and modal components are required.

4.4 Conclusions

Gettier cases have shown that a JTB is not sufficient for knowledge and that luck can undermine a true belief as knowledge. Therefore, for knowledge, an additional condition is required in addition to a JTB to ensure that a true belief is safe. There are various accounts of luck, such as the probability account, MAL and the hybrid account, but none of these accounts captures the full spectrum of luck in knowledge acquisition. After evaluating various theories of luck, I conclude that none of them can adequately explain luck. MAL has merits over other accounts of luck and can evaluate the luckiness of a single event. However, MAL fails to capture the luckiness of multiple events. Pritchard claims that veritic luck, i.e. intervening and environmental luck, undermines knowledge. Pritchard proposes ALE to ensure the safety of a true belief against knowledge-undermining luck. However, Pritchard realises that, in some cases like Temp, true beliefs are safe but that epistemic success cannot be attributed to the agent. Therefore, ALE is not an adequate theory of knowledge. Pritchard claims that ability and safety are the two master intuitions required for knowledge. Accordingly, Pritchard proposes ALVE, which captures both the agent's ability and safety in knowledge acquisition. Carter, Greco and Sosa criticise ALVE, noting that knowledge requires ability that should be suitably non-lucky, that acquiring knowledge by ability ensures that epistemic success is non-lucky and that there is no requirement for an additional codicil in ALVE to ensure safety. It is clear that luck can be beyond an agent's ability, as for Barney, and that safety cannot be ensured by the way epistemic success is achieved by the cognitive ability of the agent. Therefore, the alternative proposals by Carter and Greco fail to provide an adequate theory of knowledge. There is no additional merit in Kelp's safe-apt view compared with Pritchard's ALVE, as both require separate ability and safety conditions for knowledge. Since Pritchard's MAL has limitations, as shown in the counterexamples of Carter, Peterson and De Grefte, then ALVE based on MAL is not adequate for capturing knowledge-excluding luck. Since MAL cannot capture the full spectrum of luck, ALVE cannot be considered to be an adequate theory of knowledge. I conclude that to capture the full spectrum of luck, both probabilistic and modal components are required.

Therefore, the main objectives of this chapter, i.e. to identify an adequate theory of knowledge and to identify an adequate account of luck that can adequately capture knowledge-undermining luck especially when artefacts are involved in cognitive processes, have not been accomplished.

5 MODIFIED ANTI-RISK VIRTUE EPISTEMOLOGY

5.1 Introduction

A modal account of risk (MAR) cannot alone adequately capture either the risk associated with knowledge acquisition or the risk associated with artefacts involved in knowledge acquisition, as a modal component and a probabilistic component are both required to capture the full spectrum of risk. Such a hybrid account of risk (HAR) based on the modal and probabilistic accounts can capture the entire spectrum of risk in epistemology and in engineering. Why does engineering matter here? In engineering and other technological fields, risk in relation to an artefact is defined as the combination of the probability of the occurrence of harm and the severity of that harm. I am going to provide evidence from engineering that this combination alone cannot adequately account for the risk associated with artefacts. I am going to argue that applying a modal component can address the gaps in the current method of risk assessment in engineering, especially when there is uncertainty regarding hazards, failures and the consequences of harm, as risk cannot be fully captured by the probability of the occurrence of harm and its consequences.

This chapter focuses on the potential risks associated with a reliance on external artefacts during the production of knowledge. Further, I am going to modify Pritchard’s anti-risk virtue epistemology (ARVE) based on the dynamic relation between an epistemic agent and an artefact, as explained in Chapter 3. This chapter considers the roles of epistemic tools and the environment in knowledge acquisition and their relations with a cognitive agent. Rather than providing a set of necessary and sufficient conditions for knowledge acquisition, as in traditional epistemology, I focus on the details of these relations and identify potential pitfalls in knowledge acquisition. I propose a novel way of assessing safety in knowledge acquisition when epistemic tools have a role. I do this case by case based on the dynamic relations between the cognitive agent and both epistemic tools and the environment. This approach is aligned with risk assessments in engineering.

This chapter also focuses on the limitations of risk assessments in engineering, which are mainly based on a probabilistic assessment and cannot capture the full spectrum of risk. A modal risk assessment in engineering, as in epistemology, can address the limitations of safety assessments in engineering. In summary, a probability component is required along with a modal component to capture the full spectrum of risk in epistemology. Similarly, a modal component is required along with a probability component in engineering to capture all the risks in engineering.

In Section 5.2, I describe the limitations of Pritchard's ARVE. In Section 5.3, I discuss HAR, which utilises a modal component and a probability component to capture the entire spectrum of risks associated with the production of knowledge involving artefacts. Section 5.3 details modified ARVE (MARVE) based on the role of artefacts and the environment in knowledge acquisition and their relations with an agent. In Section 5.4, I propose a framework for MARVE that fares better than Pritchard's ARVE in evaluating epistemic risk, especially when an artefact is involved in knowledge acquisition.

5.2 Anti-Risk Virtue Epistemology and Its Limitations

5.2.1 Anti-Risk Virtue Epistemology

This section explores risk rather than luck and it considers the development of ARVE from anti-luck virtue epistemology (ALVE). As per Pritchard (2012), ALVE is:

> *S* knows that *p* if and only if *S*'s safe true belief that *p* is the product of her relevant cognitive abilities (such that her safe cognitive success is to a significant degree creditable to her cognitive agency). This proposal incorporates both an anti-luck condition (the demand that the true belief be safe) and an ability condition. (Pritchard, 2012, p. 20)

The structure of ARVE is the same, except that Pritchard prefers the concept of risk to luck. As Pritchard claims, there are merits in using risk rather than luck. Risk is more fundamental than luck. Eliminating risk eliminates luck but eliminating luck does not eliminate risk. Risk is forward looking whereas luck is backward looking. Pritchard claims that it is better to use risk than luck in the modal evaluation of the safety of a piece of knowledge. Pritchard explained the difference between luck and risk using a plane crash. Someone is lucky if they have missed the flight, which shows the backward perspective of the crash. However, risk looks forward, since taking the flight is a risk. There are differences between risk and luck. A risky event may have negative consequences. Risk is a fundamental aspect of the lack of safety.

MAR works in a similar way to the modal account of luck (MAL). Like MAL, there is an epistemic continuum where one end of the target event (failure event) is modally close, which indicates that the knowledge is at extreme risk, and the other end is modally remote and the risk is negligible. Pritchard claims that there is a continuum of epistemic risk. If the risk is modally very close, as in the lottery case, then we do not attribute knowledge because the risk of error, and thus the risk involved in having a true belief, is too high. If the risk is very far off, in contrast, then we are inclined to attribute knowledge, since there is no serious risk of error, and relatedly no luck involved in having a true belief either. Between these two extremes there is a sliding scale of epistemic risk, where, at some point – in all likelihood an indeterminate range – our inclination not to attribute knowledge stops and we start to attribute knowledge. As Pritchard argues (2020), MAL is concerned with the non-occurrence of a target event in a close possible world; however, MAR is concerned with whether the relevant risk event occurs in a close possible world.

5.2.2 Limitations of ARVE

Pritchard classifies factors in the production of knowledge as agential and non-agential. Agential factors are used in the exercise of cognitive ability. Non-agential factors include the environment. Pritchard claims that knowledge has two constraints: (1) cognitive success must be due to the manifestation of cognitive ability and (2) high-level epistemic risk must be excluded. The second constraint ensures the safety of the knowledge (i.e. protecting knowledge from extra-agential factors), such that beliefs are formed in a way that is safe, i.e. they could not easily be false. Pritchard argues that these two constraints can sometimes be in tension, since exercising cognitive ability to its maximum extent may not attain cognitive success due to extra-agential factors that have nothing to do with an agent's cognitive ability. Consider the Temp case discussed in Chapter 4. Temp's true belief about the temperature in the room is in no way connected with his ability. Rather, the hidden helper (an extra-agential factor) manipulates the temperature to make Temp's belief true. This demonstrates a key concept in Pritchard's epistemological framework: the distinction and interaction between agential and non-agential factors in the formation of knowledge. The Temp case shows that even if a belief is true and the result of a reliable process (Temp reading the thermometer), it may not constitute knowledge if it is not primarily due to the agent's cognitive abilities. Instead, the hidden helper, an extra-agential factor, plays a critical role in making Temp's belief true.

Pritchard's (2020) preferred theory of knowledge, ARVE, involves an interrelationship between the exercise of cognitive ability and potential epistemic risk due to extra-agential factors. Pritchard explains this relation in terms of epistemic dependence. In positive epistemic dependence, there is a minimal level of cognitive agency by the agent, such as in the case of Jenny discussed in Chapter 4. Jenny can acquire knowledge from a reliable informant with a minimum exercise of cognitive ability. For negative epistemic dependence, as in the case of Temp, there is a significant exercise of cognitive agency by the agent; however, an agent cannot acquire knowledge from extra-agential factors, such as the environment. Consider the case of Barney. Barney's cognitive agency, such as his perceptual abilities, is adequate for forming a true belief about a barn. However, Barney's true belief about a barn in fake barn county has a high level of epistemic risk, since in a modally close world, Barney may look at a fake barn and form a false belief about the barn. Although Barney's cognitive agency is adequate for cognitive success, an extra-agential factor such as the environment (i.e. a fake barn) makes Barney's belief unsafe. This is an example of negative epistemic dependence.

In summary, there are two key features of Pritchard's ARVE:

1. Epistemic dependence, which incorporates both positive and negative epistemic dependence due to the exercise of the cognitive agency of the agent and extra-agential factors in knowledge production.
2. MAR, which captures the safety of knowledge production. Luck is typically considered in a retrospective manner, looking back at how events occurred. Risk, as conceptualised in ARVE, can be seen as the forward-looking counterpart of luck. Risk involves considering the potential for future outcomes, particularly those that can undermine knowledge. MAR captures the safety of knowledge by assessing how likely it is for a belief to remain true across close possible worlds, thus aligning with MAL.

There are limitations to Pritchard's ARVE in terms of detailing the role and relations of agential and non-agential factors and fixing belief-forming processes in a modal evaluation of risk, especially when artefacts are involved in knowledge production. As explained earlier, ARVE "involves an interplay between manifestations of cognitive agency and extra-agential factors" (Pritchard, 2018a, p. 3066). To a certain extent, however, Pritchard accounts for the role and relations of agential and non-agential factors in knowledge acquisition when he claims that the cognitive success must at least be "significantly creditable" to ability. However, Pritchard does not give sufficient details of the role and relations of agential and extra-agential factors in knowledge acquisition. These are especially important when knowledge acquisition involves external artefacts.

As we will see in Chapter 6, Pritchard argues that if technology is properly integrated with the cognitive character of an agent, such as in extended cognition (EC), then this can result in extended knowledge (EK). Since EK involves artefacts, we must analyse the safety and risk of the role of each artefact in knowledge acquisition. The risks associated with an artefact can affect the production of knowledge. Since cognitive integration is a requirement for EK, it is of the utmost importance to know the relationship between an artefact and an epistemic agent and also the potential risk associated with the use of the artefact. It is also important to assess the reliability and availability of the artefact in cognitive processes.

To address the above gaps in Pritchard's ARVE, my strategy here is to modify ARVE based on the following:

1. Establish a suitable account of risks that can cover the entire spectrum of risks in knowledge production.
2. Establish the role of an artefact and its relationship with an agent based on the dynamic relationship between the artefact and the agent, i.e. based on the various loops detailed in Chapter 3.

Section 5.2.3 discusses HAR, which seems to be an adequate account of risk that can cover the entire spectrum of risk in knowledge production. Sections 5.2.3.1, 5.2.3.2 and 5.3 discuss: (1) knowledge acquisition involving artefacts, (2) risk assessments for artefacts and (3) the relationship between an artefact and an agent in knowledge acquisition. Once the relationship between an artefact and an epistemic agent is established for knowledge acquisition, I modify ARVE based on the interplay between the agential, artefactual and environmental factors.

5.2.3 HAR and Risk in Knowledge Acquisition

A risk assessment approach is required for EK scenarios that harmonises the cognitive dimensions (pertaining to how users engage with and interpret information from artefacts) and the technical dimensions (pertaining to the reliability and functionality of artefacts). Fundamentally, risk entails the likelihood of an adverse event that could impede objectives, goals or desired outcomes. In essence, risk revolves around the probability and impact of unfavourable occurrences. However, when EK involves artefacts, risk has a more intricate nature. We need to consider not only the conventional aspects of the probability and impact of risks but also the broader spectrum of scenarios and implications resulting from the integration of technology with cognitive processes. Such a holistic treatment of risk is pivotal for properly addressing potential negative outcomes across theoretical and practical domains. When analysing risk within the context of EK, in which cognitive processes are extended through the integration of artefacts, the notion of risk is wider:

Technical Risks: These include the risks associated with the artefact itself, such as the possibility of its malfunctioning, inaccuracy or failure, which can lead to incorrect information or decisions. A technical (engineering), risk assessment involves evaluating the likelihood of failures, errors or accidents and their potential impacts. The components of engineering risk are:

- Probability of Occurrence: This is the likelihood or chance of a particular event occurring. In many risk assessments, this probability is quantified based on historical data, statistical analysis or predictive modelling.
- Severity of Consequence: This is the impact or severity of the outcomes if the risk event occurs. It can range from minor inconveniences to significant damage, including financial loss, health and safety hazards, or environmental impacts.

Cognitive Risks: These risks arise from human interactions with artefacts, including misinterpretation of data, overreliance on automated systems and cognitive biases influenced by the use of technology.

In the context of knowledge acquisition, luck refers to the chance occurrence of acquiring true beliefs without a reliable method. In contrast, mitigating risk in knowledge acquisition involves a systematic assessment of potential errors or misinformation and then implementing strategies to reduce these risks. In line with Pritchard's veritic luck, I would like to characterise epistemic risks are veritic risks such as intervening risks the environmental risk.

Here are my objections against MAR regarding EK. Pritchard's modal approach seems to focus too much on hypothetical or possible worlds and may neglect practical, real-world risks. This overemphasis might lead to undervaluing or overlooking actual probabilistic risks that are more relevant in real-life situations. In a practical risk assessment, especially in engineering or EK contexts, real-world probabilities and a data-driven risk analysis could be more pertinent than theoretical modal considerations. The modal account may have insufficient empirical grounding. Although it theorises about possible worlds and different scenarios, it may not adequately incorporate empirical data or real-world evidence, which are crucial for a comprehensive risk assessment. This could limit its applicability in fields that rely heavily on empirical data, such as technical risks, where risk assessments must be grounded in observable and measurable phenomena. Quantifying modal risks or assessing how a system might behave in various possible scenarios may be inherently speculative and subjective. This could lead to challenges in objectively measuring and comparing risks. In technical risks, where precise and quantitative risk assessments are essential, the subjective nature of a modal risk analysis could be a significant limitation. Pritchard's modal approach might not adequately account for the complexity of some systems, especially in cases of EC involving multiple interacting components. The interplay between these components can produce emergent behaviours that are not easily predictable by considering individual elements in isolation. This could lead to an underestimation of risks for those complex systems in which emergent properties play a crucial role. There may be a challenge in appropriately balancing the attention given to modal risks (possible scenarios) and actual risks (real-world probabilities). Overemphasis on one at the expense of the other could lead to a skewed risk assessment. For practitioners in fields like engineering and EC, finding a balance between considering theoretical possibilities and focusing on practical probabilities is crucial for effective risk management.

Neither the modal distance nor probability alone cannot capture the full spectrum of risk. Moreover, two events with the same probability and the same undesired consequence can have intuitively different risk levels.

For knowledge acquisition involving an external artefact, such as NVGs and a soldier, one way of assessing the nature of risk is in terms of probability. For example, the potential for NVGs to fail can be measured using probability theory. However, such probabilities are not always aligned with our intuition about risk. For example, the probability of the failure of NVGs could be .001 based on empirical data. The probability of buying the single winning ticket from a batch of 1000 tickets is also .001. Winning the lottery (in a lottery with 1000 tickets) is modally closer in nearby worlds than the malfunctioning of NVGs, even when the probability of both is 0.001. Winning a lottery, normally not framed as a "risk" in the negative sense, is modally closer (more likely in nearby possible worlds) than the malfunctioning of a sophisticated piece of technology like NVGs. Although, winning a lottery is typically seen as a positive outcome, in the context of risk analysis, it represents a low-probability event with a significant impact. The comparison between the lottery win and the malfunctioning of the NVGs in the context of risk assessment is meant to show that a probabilistic approach alone is insufficient. It needs to be complemented with a modal analysis to capture our intuitive understanding of risk, especially in scenarios involving knowledge acquired through external artefacts. Therefore, in terms of establishing the nature and characteristics of risk, the probability of impairment of an external artefact cannot be compared with the modal component of risk. Pritchard (2005, 2015b, 2016a) makes this point by saying that the modal account and probability account are separate in every judgement. These types of events are intuitively different, although their probabilities are similar. Pritchard's strategy against using probability in a risk assessment is structurally analogous to how he argues against the probability theory of luck. Pritchard considers two events with the same probability and the same undesired consequence. However, one event is modally closer than the second. Accordingly, Pritchard concludes that the modally closer event is intuitively riskier than the other, although the probabilities of both events are the same.

Pritchard (2015b) argues that MAR is adequate for evaluating the risks associated with knowledge acquisition. However, there is evidence to show that MAR alone cannot capture the full spectrum of risk in knowledge acquisition (Carter & Peterson, 2017; De Grefte, 2020). When knowledge acquisition involves artefacts, the concept of risk is further complicated due to the risk associated with the artefact itself. In engineering and other technological fields, the risk in relation to an artefact is defined as the combination of the probability of occurrence of harm and the severity of that harm (IEC 61508, p. 24). Thus, the various risk assessment techniques in engineering are solely based on a probabilistic component. However, the risk associated with knowledge acquisition is related to veritic risk. How can these two concepts of risk be reconciled in knowledge acquisition involving an external artefact?

In engineering, a risk assessment is comprehensive, incorporating both the probability of an event and the severity of its potential impact. Note that the severity of harm is a critical component in engineering risk assessments because it directly relates to the consequences of technological failures, which can range from minor inconveniences to catastrophic outcomes. On the other hand, the traditional epistemological discussions about risk, which revolve around knowledge acquisition, have predominantly focused on the likelihood of epistemic errors or failures (e.g. acquiring false beliefs). The severity of harm is less pronounced in epistemology, possibly due to the nature of epistemic pursuits, as a "harm" is often intangible and related to the value or reliability of knowledge rather than physical consequences. Integrating a severity of harm perspective into an epistemological risk assessment could significantly enhance our understanding of the risks associated with knowledge acquisition, especially when external artefacts are involved. For instance, the incorrect functioning of a medical device due to a flawed belief could have severe consequences, just like high-severity risks in engineering. There is evidence in engineering that the combination of the probability of occurrence of harm and the severity of that harm cannot alone adequately account for the risk associated with artefacts. Historical engineering failures often provide insights into the limitations of traditional risk assessment methods. For example, consider the Space Shuttle *Challenger* disaster in 1986. Risk assessments based on probability and severity failed to foresee the catastrophic outcome. The disaster highlighted the role of organisational and communication factors that were not covered by the traditional risk models. Another instance is the Fukushima Daiichi nuclear disaster in 2011. Here, the combination of an earthquake and a tsunami led to outcomes that traditional risk assessments had failed to predict. This incident demonstrates the limitations of conventional risk models in accounting for complex, coupled natural events. In addition to probability and severity, human factors such as decision-making, communication and organisational culture play a crucial role in the manifestation of risks in engineering projects.

MAR alone also cannot adequately capture either the risk associated with knowledge acquisition or the risk associated with artefacts involved in knowledge acquisition, as both modal and probabilistic components are required to capture the full spectrum of risk. Therefore, I propose HAR. This theoretical model is designed to assess and manage risk by integrating two primary components: the probabilistic component and the modal component. This approach aims to offer a more comprehensive understanding of risk, especially in complex scenarios. Although De Grefte's hybrid account of luck and my proposed HAR both incorporate probabilistic and modal components, they are applied to different realms – luck and risk, respectively. HAR has practical applicability in real-world scenarios, especially in technology and knowledge acquisition and offers a comprehensive framework for risk assessment that is both theoretically sound and practically relevant.

Such a HAR based on the modal and probabilistic accounts can capture the entire spectrum of risk in epistemology and in engineering. The application of the modal account can address the gaps in the current method of risk assessment in engineering, especially when there is uncertainty regarding hazards, failures and the consequences of the harm, as risk cannot be fully captured by the probability of the occurrence of harm and its consequences.

HAR encompasses both a probability component and a modal component. It is an adequate theory of risk, especially for the risk in knowledge acquisition involving an external artefact, as, roughly, the modal component can address veritic risk and the probability component can address the risks associated with the artefact in knowledge acquisition. Thus, the modal and probability components in HAR can address the full spectrum of risk associated with knowledge acquisition. HAR is also adequate for risk assessments in engineering. The integration of both probability and a modal analysis in HAR addresses the limitations of traditional probability-based risk assessments by considering a wider range of possible scenarios and their implications. This approach is particularly beneficial in complex engineering systems, as the risks are not solely a function of the likelihood of events but also of their systemic interactions, context and potential impacts. By applying this comprehensive approach to knowledge acquisition involving external artefacts, HAR offers a more holistic understanding of risks, ensuring that both epistemological and engineering perspectives are adequately addressed. Thus, a hybrid account based on probability and a modal assessment would be an effective tool in addressing risk, both in epistemology and in engineering.

If an external artefact is involved in the acquisition of knowledge, how can the risk associated with the external artefact be evaluated in terms of its impact on knowledge attribution and how can the reliability of the external artefact be assessed? Risk here includes the risk of an artefact malfunctioning and the risk of an artefact functioning in such a way as to contribute to the agent’s having a false belief.

5.2.3.1 Risk Assessment with Artefacts

There are similarities and differences between the concepts of risk in engineering and in philosophy. As we have seen earlier, Pritchard (2015b) proposes MAR to address risk associated with knowledge acquisition. However, risk in engineering is solely based on a probability component with no modal component, as in philosophy. In engineering, risk is a combination of the likelihood of a hazard occurring and the severity of the outcome or consequence (Stapelberg, 2009, p. 529). However, in philosophy, there are various accounts for risk, such as the probability account, lack of control account, MAL and various hybrid accounts. HAR captures both the probabilistic component and the modal component.

In terms of knowledge attribution involving artefacts, then risk assessments and knowledge attribution are further complicated because of the relationship between an agent and an artefact and the risk associated with the artefact itself. The traditional way of assessing the risk associated with an artefact is solely through a probabilistic component. I am going to argue that in engineering, risk cannot solely be based on probability; rather, it depends on both probabilistic and modal components. Uncertainty is the core problem in risk analysis in engineering. How is uncertainty handled in engineering? The Health and Safety Executive in the UK states: "Recent studies have shown that as mankind has evolved to cope with the dangers and uncertainty of life, we have all been provided with inbuilt mechanisms for dealing with risk – mechanisms that reflect our personal preferences and the values of the society in which we live" (HSE, 2001, p. 10). There are certain elements beyond probability. The HSE (2001) guideline further states:

> The regulation of health and safety is replete with examples where the potential severity of the consequences, rather than the probability of them occurring, is the dominant consideration. This is particularly true for hazards where there is considerable uncertainty on the nature and scale of the risks they give rise to. … We therefore need to look at uncertainty in more detail.
>
> (HSE, 2001, p. 27)

The various risk assessment techniques, like hazard operability (HAZOP) and layer of protection analysis (LOPA), use a modal component to varying degrees in a disguised form in addition to probability in the consideration of safety when there is uncertainty. The risk in engineering is solely based on the probability of occurrence of harm and its consequence. However, in reality, risk is more than the probability of the occurrence of harm and its consequence, as sometimes the potential for risk events is modally closer even though the risk event is probabilistically remote. Thus, to cover the entire spectrum of safety and risk, both probabilistic and modal components are required. In a risk assessment, redundancy is used to mitigate against various possible scenarios (a modal consideration) in which a primary system might fail. This preparation is not solely based on the probability of failure but also on the consideration of different possible scenarios in which an alternative or backup system could prevent catastrophic outcomes. For example, redundancy is defined as "the existence of more than one means for performing a required function or for representing information. Examples are the use of duplicate devices and the addition of parity bits. Redundancy is used primarily to improve reliability or availability" (IEC 61508, 2010, 3.4.6). Redundancy is aligned with the principles of HAR, which combines a probabilistic risk assessment (likelihood and severity of harm) with a modal analysis (consideration of various possible scenarios and their closeness to the actual world). I discuss redundancy to demonstrate how engineering risk assessments, although predominantly probabilistic, do engage in modal thinking. This supports the proposal for HAR, which explicitly acknowledges and integrates both probabilistic and modal components to give a more comprehensive understanding of risk.

Consider a new device that should have an annual probability of failure of 10^{-3}. The designer proposes two configurations:

1. The first option is to use a single component with an annual probability of failure of 10^{-3}, i.e. the component will fail once in every 1000 years.
2. The second option has two components that do the same thing, so that there is redundancy. The annual probability of failure of the first component is 10^{-2}, i.e. it fails once every 100 years. For the redundant component, the annual probability of failure is 10^{-1}, i.e. it fails once every 10 years. The probability that both will fail in a particular year is $10^{-2} \times 10^{-1} = 10^{-3}$.

From the risk point of view, both configurations have a probability of failure 10^{-3}, i.e. once in every 1000 years. The probabilities of the occurrence of harm in both cases are the same, but the probabilities of failure of the individual components are different. There is redundancy only in the second option. From a safety perspective, which configuration is better? Although, the probability of failure of the first configuration is 10^{-3}, there is no assurance that the component will be safe all the time. In engineering and in risk assessment, “safety” typically refers to the likelihood that a system or component will not fail and cause harm or damage during its operational lifetime. It encompasses both the reliability of the component and its ability to function without leading to adverse outcomes. A system’s safety is not solely determined by the probability of failure of its components. It also depends on how the system can handle such failures, which depends on system resilience and redundancy.

The first configuration uses a single component with a low probability of failure (10^{-3}). Although statistically reliable, its safety is contingent on the single component’s uninterrupted performance. The lack of redundancy means that if this component fails, the entire system fails, thus posing a risk. The absence of redundancy means that any unexpected failure of this component directly leads to system failure. This introduces a vulnerability, despite the low probability of failure. Our subjective concept is that a failure event is modally remote – once in a blue moon – but that failure event is random and could happen today or in 10 or 100 years' time. If it happens today, that is unfortunate. However, in the second option, although the overall probability of failure of the configuration is the same as that of the first option, the annual probability of failure of either component is higher than in the first option, but the redundancy makes the system safer than the first option. In the second configuration, even if the failure event of either component is soon, the redundancy makes the system safer. However, such protection is not available in the first option. Therefore, intuitively, we can conclude that the second option is safer than the first option. Comparing cases 1 and 2, although the probabilities are same, the risk event in case 1 is modally closer than the risk event in case 2. The modal distance of failure measures how "close" or "likely" a possible failure is in the spectrum of all possible outcomes. In the first configuration, a failure (however statistically unlikely) directly results in system failure, making it modally closer in terms of impact. On the other hand, in the second configuration, the failure event is considered to be modally further away, as it requires the concurrent failure of both components for the system to fail, an occurrence less likely in the realm of all possible scenarios. This perspective provides a qualitative dimension to risk assessment that complements the quantitative probability analysis.

Here, as I argued earlier, a probabilistic component alone cannot capture the safety of the two configurations and doing so requires a modal component in addition to probability.

One risk assessment technique is HAZOP. The method entails investigating *deviations* from the design intent for a process engineering installation by a design team with expertise in different areas, such as engineering, operations, maintenance, safety and chemistry. The team is guided in a structured process, by using a set of guidewords to examine deviations from normal process conditions at various key points (nodes) throughout the process. This allows the operators to identify the *causes and consequences* of deviations.

Consider a chemical plant. It is designed is to perform under normal operating conditions, which relate to parameters such as pressure, temperature, flow rate, and the level and composition of the various streams in the equipment. The system is assessed using the guidewords, which are typically a prefix such as "low" or "high". These are applied to the parameters to give deviations from normal process conditions at key nodes, such as high pressure, low temperature, etc. A brainstorming session is used to list various causes of such deviations and the impacts and safeguards for each. For example, what can cause high pressure in the system? Is the occurrence modally remote or close (is the probability high or low)? Are there any safeguards? If not, what can be done?

Theoretically, in HAZOP, risk is still based on the product of the probability of a failure and its consequence. However, we can see how modal intuition is ingrained into this risk assessment technique. When HAZOP assesses how many changes or deviations are required from normal operation to reach a hazardous state, it is essentially making a judgement about the modal closeness or remoteness of potential failure scenarios. If a system can easily deviate into a hazardous state with minimal changes (i.e. the hazardous state is modally close), the perceived risk is higher. Conversely, if it requires multiple, unlikely changes to deviate into a hazardous state (i.e. the hazardous state is modally remote), the perceived risk is lower. A HAZOP assessment considers not just the likelihood of a specific deviation (a probabilistic view) but also the extent of deviation from normal operation required to reach a hazardous state (a modal view).

Consider another type of risk assessment, LOPA. Willey (2014) defines LOPA as a simplified risk assessment methodology used to understand how a process deviation can lead to a hazardous consequence if it is not interrupted by the successful operation of a safeguard called an independent protection layer (IPL). An IPL can prevent a scenario from progressing to an undesirable consequence. The combined effect of the IPLs associated with a hazard scenario are compared against risk tolerance criteria to determine if additional risk reduction measures are required to reach a tolerable level of risk.

Traditional risk assessments in engineering primarily focus on the probability of failure and the severity of its consequences. Probabilistic methods are excellent for quantifying known risks based on historical data or empirical evidence. However, this approach can sometimes overlook the nuances relating to system vulnerabilities when there are unknown variables, complex interactions or unprecedented conditions, which are often where the most critical risks lie. Engineering risk assessments sometimes deal with complex systems in which uncertainty is inherent. This uncertainty is not just about the likelihood of failure (probability) but also in how the system might fail (modal aspects). The integration of modal intuition with probabilistic methods in a risk assessment addresses the complexities of real-world scenarios. As discussed above, in HAZOP, for example, the team explores various “what if” scenarios and examine how slight changes in operational conditions could lead to significant risks. This process implicitly uses modal reasoning to understand potential failure modes. In LOPA, scenarios that require many simultaneous failures to occur are considered less risky (modally distant) than those in which a single failure could lead to a hazardous event (modally close). Clearly, whenever there is uncertainty, the various risk assessment techniques use modal intuition along with probability to evaluate safety. Therefore, it can be concluded that both the probabilistic and modal components are required to capture the full spectrum of safety and risks. From the above discussion, it can be concluded that probability alone cannot fully capture risk in engineering. A hybrid account can capture risk both in philosophy, especially epistemology, and in engineering. In engineering, rather than restricting risk to be the product of the probability of a failure and its consequence, it would be better to use the hybrid account, i.e. the combined probabilistic and modal accounts, to assess the full spectrum of risk. The advantage is that the hybrid account can better encompass the various types of risk assessment, such as LOPA, HAZOP, etc. Currently, engineers in practice rely on probabilities in assessing risk. However, engineers should also be factoring in modal closeness.

5.2.3.2 Application of HAR in Engineering

A risk assessment in engineering is based on probabilities and a qualitative assessment of uncertainties. Is HAR better for assessing the risk associated with artefacts in engineering compared with a traditional risk assessment using probabilities? Applying the hybrid account to engineering will provide better insights in a risk assessment. In engineering, the design intent will be met by an artefact under the right or normal conditions. If the conditions change, the artefact may not produce the desired results. By applying HAR, the risk associated with an artefact can be assessed in terms of the product of the probability of failure and its consequence. From a modal perspective, risks can be evaluated in terms of modal distance, i.e. whether the risk event is modally closer or not. Most of the time, there is a possibility that the results from the probability component match those from the modal component. However, there are cases, such as cases 1 and 2 where the probability component differs from the modal component. Case 1 is a single-component system with a probability of failure of 10^{-3}. Case 2 has two components with redundancy, both having a higher individual probability of failure of 10^{-2} and 10^{-1}, respectively.

An artefact is designed in such a way that in the real world, it functions normally and meets the design intent under normal conditions. If potential hazards can be identified by establishing a variation of the parameters from normalcy, then the modal distance can be evaluated based on how much change is required to the real world to realise the hazard in modally close worlds. If too much change is required to realise the hazard, then the event is modally distant and the risk is low. However, if little change is required, the materialisation of the hazard is modally close and the risk is higher. This type of modal assessment will provide better insights into safety and risk, especially when there are uncertainties. The significance of each change is equally important. Not all changes are equal; some may have a more profound impact on system integrity than others. For instance, a minor adjustment in operating temperature might be less significant than a change in a critical structural component. The potential impact of each change needs to be evaluated in terms of its severity and the likelihood of causing a hazard. This will be part of the probability component of HAR. Incorporating both the number and significance of changes provides a more nuanced and accurate assessment of risk. This is particularly relevant for complex systems in which uncertainties are inherent and multi-layered. Therefore, in applying HAR in engineering, it is essential to consider both the number and significance of changes from normal operating conditions. This approach gives a more comprehensive understanding of the modal dimension of risk and enhances the reliability of safety assessments in engineering practice, especially in scenarios characterised by high uncertainties.

For example, a HAZOP evaluation starts by considering the functioning of the design under normal conditions, so that in these normal conditions, the proposed design will produce the desired outcome. The guidewords in HAZOP essentially relate to the changes to the real world needed to materialise a hazard. In such cases, it would be better to evaluate whether the hazard realisation is modally close or distant based on the extent of the variations required from normalcy in the real world. If the realisation of the hazard is farfetched, i.e. modally distant, the risk is tolerable, whereas if the realisation of the hazard is modally close, the risk is intolerable.

In certain cases, norms and practices can address modally close hazards, such as in the example of life-saving rules where the risks are modally close; therefore, a cautious mindset is required for such cases to avoid the realisation of the hazards. Procedures and practices are another way to handle modally close hazards. A learner mindset allows an engineer to learn from an incident, which reinforces their ability to avoid modally close hazards.

Note that safety engineers focus on only very specific kinds of risk of failure, namely the risk of failure under something like normal conditions and possible abnormal conditions. That a certain piece of equipment is at high risk of failure because some manipulator might damage it is of little consequence to safety engineers, as this type of risk is considered negligible by safety engineers. However, for epistemologists, interference by manipulators, like jokesters, etc., has important consequences, such as in the case of Temp, where a hidden helper makes Temp's belief about the temperature true. Such cases have no importance in engineering.

In the next section, I am going to modify Pritchard's ARVE based on HAR and the dynamic relation between an epistemic agent and an artefact.

5.3 Modified Anti-risk Virtue Epistemology

In this section, I establish the nature of the relations among an artefact, the environment and an epistemic agent and also the criteria for what is to be fixed in an initial belief-forming process for a meaningful modal risk assessment. As discussed earlier, the overall aim of this chapter is to address the limitations of ARVE based on HAR, which has probabilistic and modal components and clearly establishes the role of external artefacts and the environment and their relationships with a cognitive agent. By including both probabilistic and modal components, HAR can assess the risk associated with external artefacts more effectively than ARVE. It considers not just the probability of an artefact failing but also the various conditions under which this failure might occur. It applies the modal distance, which helps in understanding the likelihood of an artefact failing under slightly different conditions. This approach acknowledges the interplay between an agent's cognitive abilities, the tools they use and the environment in which they operate, thus offering a more holistic view of the knowledge-acquisition process. The main consideration of this modification is knowledge acquisition involving artefacts.

The important difference in MARVE with respect to Pritchard's ARVE is that the new version establishes the relationship between an epistemic tool and a cognitive agent and clearly defines the interfaces among environmental factors, cognitive processes and risk events. Based on the relations among an artefact, the environment and an epistemic agent, MARVE explores what is to be fixed as an initial belief-forming process in a modal risk assessment.

The objective of Sections 5.3.1 to 5.3.5 is not to establish anti-Gettier safety conditions and necessity and sufficiency conditions for knowledge. Rather, the focus is to analyse the relationship between an artefact and an agent and to show how the agent's relationship with the environment is different from their relationship with an artefact. This analysis will provide more insight into individuating the belief-forming methods and fixing the relevant initial conditions for a modal assessment of risks that excludes luck.

5.3.1 Relationship Between an Artefact and an Agent

In Chapter 3, I describe the relationship between an artefact and a cognitive agent based on various loops, such as open loops, feedback loops, feedforward loops and emulation loops. There is the potential that a feedback loop will not constitute a reliable way of forming beliefs. However, our discussion here is limited to the feedback loop between a cognitive agent and an artefact that can reliably produce true beliefs.

My argument in Chapter 3 is that a feedback loop with an external artefact and the manipulation of that artefact to achieve cognitive success are necessary and sufficient for EC, i.e. for extending cognitive processes into the environment beyond the brain and body of the cognitive agent. However, as discussed in Chapter 3, with dynamic progress, such a feedback loop can become the basis for an acquired skill or a developed cognitive pattern, which itself can then become the basis for further open and feedforward loops during cognition.

In some cases, cognitive agents work together to achieve joint cognitive success. In other words, cognition is distributed among several people and each individual participates in the cognitive task, for example in a scientific project in which all the scientists are working together to achieve the common objectives, resulting in cognitive success. In such cases, many artefacts and agents will be involved in cognition. How can the risks associated with such artefacts be evaluated?

The extent of the risk associated with an artefact in knowledge attribution varies. There are two types of risk relating to an artefact in knowledge acquisition:

(1) The risk that the artefact will malfunction and fail to provide any practical benefit when used. An example is the failure of NVGs.

(2) The risk that a subject relying on the artefact will form a false belief due to information provided by the artefact. An example is the environmental risk discussed earlier where a soldier formed a true belief about a true missile among

> fake missiles in an enemy’s arms depot, which is epistemically fragile, as in modally close worlds, she could be looking at a fake missile and form a false belief about the missile.

The integration of NVGs into the soldier’s cognitive process extends beyond mere tool use; it alters the way information is processed and interpreted. Both malfunctions and misinterpretation risks highlight the importance of understanding how external artefacts integrate with and influence cognitive processes. HAR, with its dual focus on probability and modal closeness, provides a more comprehensive framework for assessing these risks. It considers not just the likelihood of NVGs failing but also how slight changes in circumstances could significantly impact the soldier’s decision-making process.

Note that (1) and (2) might not always be the same. The first depends on the probability of failure of the artefact in knowledge production, whereas the second relates to the potential of veritic luck in knowledge production.

Pritchard classifies factors in the production of knowledge as agential and non-agential. Agential factors are used in the exercise of cognitive ability. In the above example, the soldier’s perceptual abilities and skill in using the NVGs are agential factors. A variation in the non-agential factors, such as the environment, can easily prevent the acquisition of knowledge, such as a true missile among fake missiles in an enemy’s arms depot. Faulty artefacts can increase the risk in knowledge acquisition. For example, a faulty thermometer or clock can result in a lucky true belief if, by mere coincidence, the data provided by the faulty clock or thermometer matches the actual time or temperature at the moment of perception. In this case, the risk is associated with the failure of a single artefact. However, in large scientific projects, such as the Large Hadron Collider, many artefacts and many agents are involved in the cognitive success and knowledge attribution. So, how are the risks to be evaluated in such projects? The role of epistemic tools in cognition is versatile and important. To capture factors in the production of knowledge completely, the role of such tools in cognitive processes must be specified. The failure of epistemic tools, such as a broken thermometer, shows that epistemic tools affect the production of knowledge.

Therefore, rather than just agential and extra-agential factors, there are three types of factors in the production of knowledge: (1) agential factors, (2) factors relating to epistemic tools and (3) environmental factors. My approach in distinguishing agential, epistemic tool-related and environmental factors aims to provide a more nuanced understanding of knowledge production, especially in contexts involving EC. EC postulates that external artefacts can become integral parts of our cognitive processes. Therefore, it is important to identify and differentiate the roles played by these different factors in knowledge production. The integration of artefacts in cognitive processes raises critical questions about their role in knowledge production. If artefacts are part of the cognitive system, their reliability, functionality and integration directly impact the epistemic validity of the knowledge produced. Thus, it is necessary to expand the epistemological inquiry to include external tools and their interaction with the agent. Due to EC, categorising the factors involved in knowledge production into agential, epistemic tool-related and environmental helps in understanding how external artefacts, when integrated into our cognitive systems, can constitute part of our knowledge-producing processes. This approach aligns with the EC framework by acknowledging the active role of external artefacts in cognitive processes and their impact on epistemology.

Agential factors include cognitive faculties such as memory, reasoning, introspection, perception, cognitive traits, cognitive character and disposition. In terms of cognitive processes and cognitive success, there is a significant difference in the interface between an epistemic tool and the agent compared with the interface between the environment and the agent.

5.3.2 Interface Between Epistemic Tools and a Cognitive Agent

An epistemic tool is an artefact, instrument or technological object that is an aid to cognition or has a constitutive role in cognition. As discussed in Chapter 3, reading glasses are an aid to cognition, but NVGs properly integrated with a soldier are an example of EC based on feedback loops, as the goggles play a constitutive role in the cognitive processes. The use of an epistemic tool and its integration with the agent in cognitive processes normally create the potential for epistemic action. In EC, cognitive faculties and epistemic tools are together required to achieve cognitive success.

The interface between epistemic tools and cognitive faculties can be an open loop, as for Temp’s broken thermometer. Consider that Temp’s thermometer begins to function normally. Temp does not question the reliability of the instrument and just measures the temperature. This is an example of an open loop, since the thermometer is aiding Temp’s cognition to achieve cognitive success in measuring the temperature. However, if Temp understands how a thermometer works and its potential failure modes, such as the need for a power supply, he would check the reliability of the instrument before measuring the temperature. Note that different levels of cognitive integration are possible with the same artefact and agent. For example, the relation between the NVGs and the soldier could be an open loop or a feedback loop depending on how the soldier interacts with the NVGs and how the NVGs are cognitively integrated with the soldier. If the NVGs are automatically set and only the perceptual abilities of the soldier are required to form beliefs using them, then only an open loop exists. However, if the soldier manipulates the NVGs and the goggles are cognitively integrated with the agent to achieve cognitive success, then there is a feedback loop. Thus, various loops, such as an open loop or a feedback loop, are possible with the same artefact depending on how the agent interacts with it. For example, I know that my watch requires a battery, but I don’t always check the reliability of my watch before forming a belief about the time. This case does not have a feedback loop. Rather, there is an open loop. However, I suggest that there is a feedback loop between NVGs and soldier when there is a two-way interaction between them that results in continuous reciprocal causation (CRC). In this case, the NVGs have become integrated with the soldier and the feedback loop has a constitutive role in cognition, rather than being an aid to the cognitive process.

A feedback loop has the potential to produce an epistemic action. The purpose of an epistemic action is to achieve cognitive success. Therefore, the conclusion is that the same epistemic tool can have either a constitutive role or a mere helping or enabling role in a cognitive process, depending on the context. As mentioned earlier, cognitive faculties and epistemic tools can be integrated to achieve cognitive success.

5.3.3 Interface Between the Environment and a Cognitive Agent

The interface between the environment and a cogniser is significantly different from the interface between an epistemic tool and the cogniser. An unsuitable environment can result in cognitive failure, irrespective of the exercise of cognitive abilities by the agent. In a risk assessment that takes into account the environment, the modal component allows us to assess whether the cognitive failure event is modally close. In contrast, in a risk assessment of an epistemic tool, the probability component is used to see whether the cognitive failure is remote or close.

As we have seen for Barney in fake barn country, environmental luck can result in cognitive failure in a modally close world and undermine the acquisition of knowledge. Gettier-style cases, such as the intervening luck for Roddy, are due to environmental interference, which is beyond the faculties of the agent’s cognitive function. A Gettieriser, such as a malevolent or benevolent demon, can be part of the environment and can affect cognitive success. Therefore, we can conclude that the role of the environment in knowledge acquisition is beyond the limits of the cognitive agent. The environment's impact on cognitive success is more complex, and the modal component becomes crucial here. It assesses how modally close or far a potential cognitive failure event is by considering variations to the environment. The key point is that environmental factors affect cognitive success in ways that are not directly proportional to their likelihood or frequency, unlike the probability of tool failure. Environmental impacts are assessed based on their potential to change the scenario in nearby possible worlds, not just on their probability in the actual world. Unlike the predictable nature of tool failure, environmental factors can vary greatly and unpredictably; thus, they affect cognitive tasks in non-linear ways. This complexity is what makes the environment's role in cognitive processes distinct from that of epistemic tools.

Epistemic risks are associated with the factors of cognition, such as faults with cognitive faculties, an inappropriate environment or faulty epistemic tools. A failure event of one of these factors of cognition can result in undesired consequences, such as a false belief. Figure 5.1 illustrates examples of the factors in cognition, possible failure events and the undesired consequences.

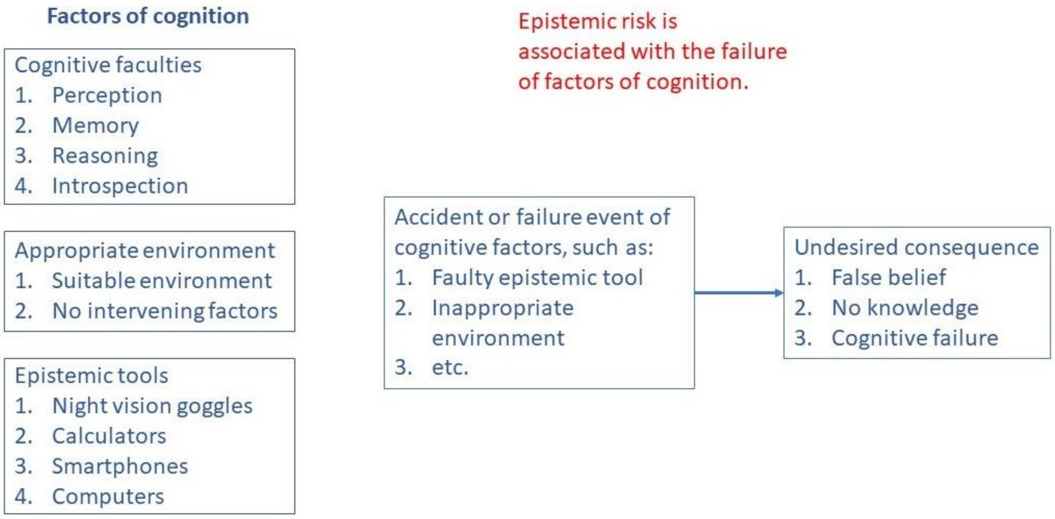


Figure 5.1. Epistemic risk: The interface between factors of cognition, cognitive failure and consequence.

Figure 5.1 shows the epistemic risk in relation to the factors of cognition, which may fail and result in an undesired consequence, i.e. cognitive failure. The left-hand boxes show the factors of cognition, such as cognitive faculties and epistemic tools. In addition to cognitive faculties and epistemic tools, an appropriate environment plays a key role in cognition and cognitive success. The box in the centre shows the potential epistemic risks, i.e. the risk event and the reasons for such risks. A risk is associated with the magnitude of the failure or deviation of the cognitive factors. The reasons for such occurrences are due to:

(1) Accidental failure of cognitive agency: The failure of the cognitive agency of the agent, for example, blurred vision, can result in false beliefs. Any impairment in cognitive faculties, such as perception, memory, reasoning, introspection, etc., can cause false beliefs. The risk can occur due to the epistemic ignorance of the agent. For cognitive success, here I am considering that the agent is competent and reliable enough to secure the agential part of knowledge acquisition. This means that the agent's cognitive faculties are functional and not impaired and that the agent is competent to handle an external artefact, if it performs normally.

(2) A risk event due to a variation of the interface with the environment: A fake environment increases the epistemic risks, such as a true missile among fake missiles or a true barn in fake barn county. A true belief formed in such an environment may be epistemically fragile.

(3) A failure or deviation event associated with a failure of the cognitive faculties or epistemic tools: As discussed earlier, unreliable artefacts, like a broken thermometer or faulty NVGs, can pose a risk in knowledge acquisition.

The right-hand box shows some of the consequences of such epistemic risks: (1) false belief, (2) no knowledge, (3) cognitive failure.

5.3.4 Environment and Cognitive Processes in a Modal World

As mentioned earlier, there are two aspects of HAR: (1) the probability of a risk event and (2) the modal distance of the risk event from the actual world. Generally, for cognitive agents, there is clearly a linear relationship between the probability of risk and the modal distance of risk. Normally, the probabilistic and modal distances for a cognitive failure event of cognitive agencies are linear in the actual world as well as in modally close worlds, just like a misprint of a lottery result in a newspaper is probabilistically and modally unlikely. The occurrence of a misprint in the lottery results is considered to have a low probability because newspapers typically have multiple checks and safeguards to ensure accuracy in reporting, especially for something as significant as lottery results. In modal terms, in most nearby worlds, the systems and checks in place would work as intended to ensure the correct publication of the lottery results. Thus, a misprint of a lottery result in a newspaper is probabilistically and modally unlikely. However, as seen earlier for NVGs, the functional failure of the goggles is linear (probability component), but due to variations in the interface between the environment and the epistemic agent (modal component), environmental factors can vary greatly and unpredictably and affect cognitive tasks in non-linear ways.

The interface between the environment and cognitive process raises an interesting question about what is to be fixed in a modal assessment. The initial conditions are fixed, as is the belief-forming process. What can vary in the modal assessment is only the interface between the environment and the cognitive processes, like for the fake barn façades. When forming a belief about the barn, Barney is looking at a true barn in the actual world, but in a modally close world, Barney may be looking at a fake barn. In this case, in the modally close world, only the interface between the environment and the belief-forming process has changed.

5.3.5 MARVE: Risk Assessment in Knowledge Acquisition

We have already discussed HAR, risk assessments with artefacts as well as the relations among an artefact, the environment and an epistemic agent. Note that the important difference in MARVE with respect to Pritchard’s proposal is that the new version establishes the relationship between an epistemic tool and a cognitive agent and clearly defines the interfaces between environmental factors, cognitive processes and risk events in a modally close world.

5.3.6 What Is to Be Fixed in a Modal Assessment?

In this section, I am going to evaluate what is to be initially fixed for a sensible modal assessment. Based on our previous discussions in Sections 5.3.1 and 5.3.2, I am going to explain how a risk assessment can be done in MARVE.

The aetiological function is to be fixed along with the initial belief-forming process. Peter Graham (2012) proposes conditions for epistemic entitlement to ensure the safety of the acquired knowledge using reliability as an aetiological function of the cognitive faculties. The aetiological function ensures that there is a suitable environment for the reliable cognitive processes in the acquisition of knowledge. Mona (2016) argues that an aetiological function achieved via a history of positive biological feedback is neither necessary nor sufficient for epistemic entitlement. However, my objective here is not to explore any relation between an aetiological function and epistemic entitlement. Rather, my purpose is to utilise the concept of an aetiological function to fix the relevant initial conditions in a modal assessment. As Graham (2012) stated, "The notion of aetiological function covers all sorts of entities, including some artifacts and learned behaviors" (Graham, 2012, p. 457). However, Graham's (2012) focus was on an aetiological function for biological kinds. I prefer to extend the aetiological function to encompass both biological kinds and artefacts. What is an aetiological function? An aetiological function is a normal function of an item in normal conditions. Reliability is an aetiological function relevant for cognitive success. Reliability is applicable to both cognitive faculties and epistemic tools. Essentially, aetiology is about attributing functions (in the sense of purposes) by virtue of the effects of a particular trait. Thus, attributing a function to a trait is a matter of pointing to the effects that account for why the trait has been selected. To determine what the function of an item is, according to the aetiological account, we must consider the effects of that item in normal conditions. If a trait is selected naturally because of its effects in normal conditions, it may not fulfil its function (or purpose) if it is not in its normal conditions. In this case, the item does not lack the ability to function, just that it cannot fulfil that function because it is not in its normal conditions. For our perceptual apparatus, our sense organs and their related belief-forming processes were selected because their effect in normal conditions was that of promoting true beliefs and avoiding false ones.

How does an aetiological function help a modal risk assessment? Consider Barney. In fake barn county, Barney's belief about the real barn is unsafe, as in a modally close world, Barney may be looking at a fake barn and form a false belief about the barn. Here, along with the initial belief-forming processes, the aetiological function is fixed, i.e. Barney's perceptual apparatus is functioning normally in a normal environment. In a sensible modal assessment, we cannot assume that, in a modally close world, Barney's perceptual apparatus is impaired, for example, because Barney is intoxicated or because a bio-jokester impaired his perception. The examples seem to be trivial. However, I am emphasising this because, in a modal risk assessment, we must consider the aetiological functions of the organism and artefact to be given.

There are three aspects to a risk assessment in knowledge production:

1. Aetiological functions: These are fixed as initial conditions for the belief-forming process and are kept fixed to allow a meaningful modal assessment of risk.
2. Risk assessment of artefacts: There is a linear relation between the probability of the risk associated with a failure mechanism of an artefact and the modal distance of the risk.
3. Modal assessment: This relates to the risks associated with a change to the environment in close possible worlds.

Epistemic risk can best be defined as the expectation of a risk event, e.g. due to problems with the acquisition of knowledge arising from a cognitive failure of cognitive factors such as (1) cognitive faculties (perception, memory, reasoning, introspection, etc.), (2) epistemic tools (NVGs, smartphones, etc.) and (3) the environment.

For example, consider a cognitive success in the actual world. How can we ensure such a cognitive success is safe and not a failure in a modally close world? For a cognitive success in the actual world, a risk event, i.e. a cognitive failure in a modally close world, may result in an undesired outcome, such as no knowledge, a false belief or a shortfall of knowledge. A deviation of a cognitive factor has the potential to form such an undesired outcome, which undermines knowledge acquisition. By looking closely at the factors of cognition and the potential deviations, it can be seen that the possible impact of a deviation of a cognitive factor has a very different impact on the overall acquisition of safe knowledge.

The distinction between the agent–tool and agent–environment interfaces is critical in understanding how cognitive processes interact with external elements. When we consider an artefact, especially in the context of EC, we are referring to an artefact that is cognitively integrated with the agent. This integration implies that there is a more direct, often intentional, interaction between the agent and the artefact, which impacts cognitive processes in a specific, functional manner. On the other hand, the environment encompasses a broader range of external factors that may not be directly integrated into the agent's cognitive processes but can still influence them. The environment includes not only the physical surroundings but also social, cultural and situational factors that can affect cognitive outcomes. Cognitive extension is relevant here because it highlights the role of external artefacts in cognitive processes. If an artefact is cognitively integrated (e.g. a soldier using NVGs), it becomes an extension of the agent's cognitive system. This integration can alter the way cognitive tasks are performed and how knowledge is acquired. It is important not just to account for modality and probabilities when an artefact is involved but also to understand how the integration of that tool changes or extends the cognitive capacities of the agent.

HAR is particularly important in the context of EK due to the complex nature of the interaction between an agent and external artefacts that are integral to the knowledge-acquisition process. In EC scenarios, these external artefacts are not merely tools used by an agent but have become part of the cognitive process itself. EK scenarios involve both modal risks (related to different possible scenarios) and probabilistic risks (related to likelihood or chance). The artefacts can fail (probabilistic risk) or may not operate as expected in different scenarios (modal risk).

In summary, differentiating between the agent–tool and agent–environment interfaces are not about disregarding the importance of cognitive extension but rather about recognising the distinct ways in which tools and environmental factors interact with cognitive processes. Understanding these interactions is crucial for accurately assessing risks in cognitive tasks, especially when external artefacts are involved. In scenarios with enabling artefacts, such as in non-EK acquisition, the artefact’s role is more straightforward, and the modal aspects of risk (how different possible scenarios might impact knowledge acquisition) are typically less complex.

As described earlier, the interface between the environment and cognitive processes can have a non-linear and unexpected effect on cognitive success that is beyond the abilities of the agent. Therefore, the interface between the environment and cognition must be segregated from the interface between an epistemic tool and an agent. The probability component provides insights about the risk.

5.4 MARVE Framework

The focus of this section is to develop a framework for MARVE. As we have seen, Pritchard’s ARVE has limitations, as it does not clearly define the role of an artefact or the environment during knowledge acquisition or the dynamic relation between them and the cognitive agent. Moreover, Pritchard’s MAR cannot capture the full spectrum of risk, especially when an artefact is involved in knowledge acquisition. In summary:

1. The above discussions show that HAR is suitable for capturing the full spectrum of risk in knowledge acquisition.

2. Moreover, these discussions clearly explained the role of artefacts and the role of the environment and their dynamic relations with an agent.

Now, the strategy is, with these resources, to see how we can overcome the limitations of ARVE. How can ARVE be modified to accommodate extended cognition involving artefacts? How can we address the risk associated with extended cognitive processes that result in knowledge acquisition?

In line with Pritchard's ARVE, for MARVE also, the agent's true belief should arise from the ability of the agent. That is, the true belief formed should be significantly credited to the ability of the agent and it should not be epistemically fragile. Moreover, the epistemic risks associated with knowledge acquisition should be minimal when attributing knowledge. Therefore, the core, i.e. the theory of knowledge, is the same as ARVE. The difference lies with how the epistemic risks are evaluated in terms of the hybrid account of luck when an artefact is involved in knowledge acquisition. As explained earlier, the purpose behind modifying ARVE is not to provide a set of necessary and sufficient conditions but to explain the roles of an agent and artefacts in knowledge acquisition and how epistemic risks can be evaluated for such a configuration. This is especially relevant in naturalised epistemology. Consider a scientific project with various systems and subsystems involving external artefacts. How does a scientist form true beliefs by utilising such complex artefacts? How does MARVE work for such a configuration? How are true beliefs formed in the actual world? What are all the epistemic risks in a modal world? When can knowledge be attributed to the agent's true belief?

The following generic framework is proposed for MARVE when an external artefact is involved in knowledge acquisition. Consider an agent forming a true belief in the actual world such that the cognitive success involves an exercise of significant cognitive ability by the agent and the use of an external artefact. The following framework applies for MARVE.

First, we need to identify the factors in cognition. The following three aspects need to be considered:

1. **Agential Factors:** Identify the agential factors in the cognitive achievement, i.e. the exercise of cognitive agency in cognition, such as the use of acquired skills, memory, reasoning, introspection, etc.

2. **Epistemic Tools:** Identify the role of external artefacts in cognition and their relation with the agent. Establish the type of dynamic relation in knowledge acquisition, i.e. whether there is a loop-based relation between an external artefact and the agent. As discussed in Chapter 3, both a feedback loop and the manipulation of an external artefact are required for extended cognitive processes. The distinction between open and feedback loops is crucial in the MARVE framework, especially when evaluating the role of external artefacts in cognition. This distinction affects how we understand the interaction between an agent and an artefact, and consequently, how we assess the reliability, effectiveness and risks associated with the artefact in the process of knowledge acquisition. If there are feedback loops, the artefact becomes a more active participant in the cognitive process, potentially influencing the agent’s belief-formation and decision-making processes. Open loops involve a one-way interaction in which the agent’s actions do not receive direct, immediate feedback from the environment or artefact.
3. **Environment:** Assess whether the agential factors and the epistemic tool are functioning normally in a normal environment. Any potential for a variation in the environmental factors and their impact on the exercise of cognitive agency by the agent and the use of an epistemic artefact in the environment are to be identified.

Second, we apply HAR to identify any potential epistemic risk in knowledge acquisition. The following aspects are to be considered. Before applying modal components to identify potential epistemic risks, the potential risks in the actual world are to be evaluated, such as the probability of failure of an artefact, etc. The first aspect in a risk evaluation is to identify potential risks with an external artefact in the actual world. The probability component is to be used for this assessment. After that, the initial belief-forming process is to be fixed in the modal assessment. What factors in the initial belief-forming process must be fixed?

1. **Agential factors to be fixed:** If an agent uses their clear memory in the actual world to form a true belief, then in each close possible world, we have to consider that the agent exercises the same cognitive ability, i.e. they exercise their clear memory in each close possible world. A sensible evaluation of epistemic risk is not possible if we consider that a bio-jokester impairs the memory of the agent so that the agent forms a false belief in a close possible world.

2. **The performance of epistemic tools and their dynamic relations with the agent to be fixed:** The performance of each epistemic tool is to be fixed to what it is in the actual world. Allowing these aspects to change would complicate the modal analysis significantly. If both the agent's cognitive abilities and the performance of tools could vary across possible worlds, it would become exceedingly difficult to predict or assess the likelihood of cognitive success or failure. By keeping agential factors (like memory) and the performance of epistemic tools (like a thermometer) consistent across possible worlds, we ensure that there is a stable basis for assessing cognitive success. This stability is important because it isolates environmental factors as the primary variables in modal assessments. If agential factors or tool performance varied across worlds, that would introduce too many variables and make it difficult to assess the specific impact of environmental changes on cognitive success. If a thermometer works properly in the actual world so that the agent forms a true belief, then we have to consider that the thermometer works properly in each close possible world. We cannot assume that a normal thermometer works in the actual world but is broken in a close possible world. If an external artefact has a feedback loop with the agent in the actual world, then that too is to be fixed in each close possible world. We cannot assume that a loop has changed in a close possible world.

What can vary in a close possible word is the environment. This includes interference with the environment so that the exercise of an agent's cognitive ability or the use of an epistemic tool is affected, resulting in a false belief. Because of a change in the environment, whatever the exercise of cognitive ability by the agent or their use of an artefact, the agent may fail to achieve a true belief in a close possible world.

Third, we verify whether the epistemic risks are tolerable when attributing knowledge. Once a probabilistic risk assessment for each artefact is complete and a modal assessment of the exercise of agential factors and the use of epistemic tools along with any potential variation in the environment has been done for each close possible world, then the next step is to evaluate whether the epistemic risks are tolerable when attributing knowledge. If the epistemic risks are remote, i.e. the agent forms a true belief in all close possible worlds, then we can attribute knowledge since the true belief formed by the agent in the actual world is not epistemically fragile in close possible worlds. However, if the epistemic risks are significant for the true belief formed by the agent, then we cannot attribute knowledge. In that case, the true belief formed by the agent is epistemically fragile in close possible worlds, i.e. the agent has formed a false belief in at least one close possible world.

We will now consider an example to show how MARVE works. We consider the previous example of a scientific project with a scientist conducting experiments. As discussed earlier, the factors in cognition include cognitive faculties, cognitively integrated epistemic tools and the environment. The scientist may rely on her cognitive faculties (such as learned skills, analytical abilities, perception, memory, reasoning, etc.) and use a systematic scientific approach (involving experiments, observations, analysis, conclusions, etc.) to form a belief. The experimental set-up may include various systems with various artefacts and pieces of equipment. Some of the external artefacts may be in a feedback loop with the scientist, which may play a significant role in the scientist's belief-forming processes.

Now, consider that the scientist forms a true belief in the actual world. How can we identify the epistemic risk associated with such belief formation? We can apply HAR to evaluate the epistemic risk and see whether the true belief formed by the scientist can be considered as knowledge or whether the belief is epistemically fragile.

Initial set-up and agential factors: The scientist forms a true belief in the actual world. To assess the epistemic risk associated with this belief formation, we first consider the scientist's cognitive abilities and skills. These agential factors are assumed to be constant across all possible worlds in the analysis.

Assessment of an experimental system using probability: Before going into the modal assessment, we evaluate the reliability of the experimental set-up using the probability component of HAR. This step involves assessing the likelihood of failure in the experimental system and determining whether this potential for failure is sufficiently remote. Essentially, we are asking: How likely is it that the experimental set-up will fail and potentially lead to false beliefs?

Application of the modal component in HAR: After establishing the reliability of the experimental system, we apply the modal component. This involves varying the environmental factors while keeping the agential factors and the performance of the experimental set-up constant. We explore different possible worlds close to the actual world to see if there are scenarios where the scientist might form a false belief.

Evaluating epistemic fragility: The key question here is whether the true belief formed by the scientist in the actual world is epistemically fragile. In other words, is there at least one close possible world where the scientist would form a false belief due to changes in the environment, even though the agential factors and the performance of the experimental system remain the same?

Once all the risks have been identified, we verify whether the epistemic risks are too high for knowledge to be attributed. If the epistemic risks are low enough, the true belief formed by the scientist in the actual world is knowledge. Otherwise, if the risk is too high, then the true belief formed in the actual world cannot be considered as knowledge as the scientist formed a false belief in a close possible world.

5.4.1 How MARVE Fares Better Than Pritchard's ARVE

The traditional way to evaluate knowledge does not particularly give any importance to the potential risk due to an external artefact in knowledge acquisition. However, as seen in Chapter 3, in some cases external artefacts play a key role in extending cognitive processes into the environment beyond the cognitive agent's brain and body. In the forthcoming Chapter 6, I discuss some of the extended cognitive processes that are candidate sources for EK. In such cases, the external artefact plays a constitutive role. Pritchard's ARVE does not have a tool to assess the risk due to the external artefact. MAR is adequate for identifying risks with veritic luck, such as environmental luck as in the case of Barney and intervening luck as in the case of Roddy. For traditional cases with veritic epistemic luck, such as Barney and Roddy, both MARVE and ARVE provide the same results because MARVE has a modal component as in ARVE. So, what is the difference between ARVE and MARVE? The difference arises when an external artefact is involved in knowledge acquisition, as in EC and EK.

Consider Pritchard's (2018a) example of NVGs and the soldier, as discussed in Chapter 3. When the soldier uses the NVGs for the first time, a feedback loop is set up (what she sees will guide where she looks, so that she learns to adjust the settings of the device to suit her preferences, and so on). However, over time, the soldier will become familiar with the instrument and will completely integrate it into her overall cognitive processes in a seamless fashion, so that she forms beliefs unreflectively.

I argue in Chapter 3 that a soldier using NVGs can be modelled using Beer's (2000) framework and that this application of dynamical systems theory to cognition entails EC because the cognition and the behaviour are inseparable in such an integrated system with feedback loops. Manipulating such tools via a feedback loop enables one to accomplish cognitive tasks that could not otherwise be fulfilled via normal open loops. Pritchard considers that this example of EC is a suitable candidate for EK. I agree with Pritchard. However, I do not go into the details of EK here, which will be discussed in Chapter 6. The issue here is how Pritchard's ARVE can ensure the safety of the target belief formed by the soldier. As per Pritchard's modal account, the soldier forms a true belief in the actual world via extended cognitive processes using the NVGs. In modally close possible worlds, the soldier will also form true beliefs using the NVGs as there is no in change in the environment. Pritchard's modal account is adequate for capturing veritic luck. However, in this example, the NVGs play a constitutive role in cognitive processes. So, how can the risk associated with the artefact be evaluated? The modal account considers only knowledge-excluding veritic luck, such as environmental and intervening luck, and it has no tools to evaluate the risk associated with the NVGs. MARVE and HAR provide a solution to this issue. As described earlier, HAR has two components, a modal component to guard against veritic luck and a probability component to evaluate the risk associated with an artefact. Because of HAR, MARVE can rightly assess the risks with the NVGs. If the goggles are brand new from a reliable manufacturer, the potential for failure may be 10^{-3} and thus, remote. Therefore, true beliefs formed via the goggles are safe. However, if the goggles are old and unreliable, the potential failure mode may occur once in 10 years, so that the potential for forming a false belief via the goggles is near. If the goggles are old or unreliable, the modal proximity of a malfunction and, thus, the formation of a false belief is much higher. In such an assessment, the complete spectrum of risk associated with the candidate for EK can be evaluated.

This risk assessment is not aligned with traditional epistemology. However, considering the potential for EC and possible EK, such a framework for risk assessment is relevant. It also provides a way to naturalise epistemology. In a larger picture, humanity has acquired and accumulated valuable knowledge through the use of artefacts. However, there is no way to evaluate the risks due to an epistemic artefact in knowledge acquisition. In this respect, MARVE is better than ARVE. Moreover, MARVE provides better insights into what is to be fixed in modal, possible worlds.

5.5 Conclusions

Pritchard's MAR has limitations as it cannot capture the full spectrum of risk in knowledge production, especially when an artefact is involved, as the risks associated with an artefact in knowledge production require both probabilistic and modal components. A hybrid account of luck covers the full spectrum of risk in knowledge production, as it includes both the modal and probabilistic accounts in a risk assessment.

Pritchard's ARVE has limitations in terms of establishing and differentiating the relation and interface between an artefact and an agent and the relation and interface between the environment and an agent. This necessitates a requirement to modify ARVE, which I did by establishing the nature and relationship of the interface between an artefact and an agent and the interface between the environment and an agent. MARVE integrates feedback loops and artefact manipulation from EC. These elements are crucial in understanding how artefacts contribute to cognitive processes and knowledge acquisition. This integration allows MARVE to evaluate the risks associated with artefacts in a more nuanced manner that reflects their role in extended cognitive systems.

HAR has probabilistic and modal components that can capture the full spectrum of risks. Thus, HAR can be applied to epistemology and to engineering. This is especially important when an artefact is involved in knowledge acquisition. In engineering, risk is solely based on probability; however, a modal component can capture the risk and safety, especially when there is uncertainty in the risk assessment.

6 EXTENDED KNOWLEDGE

6.1 Introduction

In Chapters 2 and 3, I discussed the hypothesis of extended cognition (HEC) and the extended mind (EM) within the philosophy of mind. HEC claims that cognitive processes extend beyond the skin bag of the cognitive agent. The objective of this chapter is to consider the ramifications of HEC in epistemology.

This chapter focuses on the implications of HEC for ALVE as well as criticisms, issues and various debates relating to extended knowledge (EK). In line with the criticisms raised against HEC, Aizawa (2018) raises criticisms against EK, such as the coupling-constitution fallacy and cognitive bloat. To address the criticisms and concerns raised about the possible ramifications of HEC in epistemology, I am going to formulate modified extended knowledge (MEK) based on modified ARVE (MARVE), which was detailed in Chapter 5. MEK describes the factors involved when an external artefact is involved in knowledge acquisition. It also identifies the relevant initial factors that are to be fixed in a modal safety assessment.

In Section 6.2, I explain Pritchard's attempt to formulate EK by assimilating HEC with his preferred theory of knowledge, ALVE. Section 6.3 details the requirement of cognitive integration needed for an extended cognitive process to become a candidate for EK. This section covers Carter and Kallestrup's (2020) and Palermos's (2014b) arguments for cognitive integration and potential criticisms against their arguments. To address the issues with the formulation of EK and cognitive integration, in Section 6.3.1, I propose MEK based on modified EM, which was discussed in Chapter 3, and MARVE, which was discussed in Chapter 5.

Section 6.3.2.1 details the debate between Vaesen (2013) and Kelp (2014). I conclude that the example of extended cognition (EC) proposed by Vaesen is not a candidate for EC as it does not meet the requirement for EC discussed in Chapter 3. In Section 6.3.2.3, I describe the issues that Kelp proposes against EK and modal safety assessments. After a critical evaluation, I conclude that Kelp's EK case is not aligned with the standard formulation of EC and that Kelp's argument against a modal assessment can be resolved by ensuring that the relevant initial conditions are fixed in a modal assessment, as detailed in Chapter 5. Section 6.3.2.3 presents the concerns raised by Carter (2013b) when assimilating EC in epistemology, especially those based on the nature and characteristics of epistemic luck. After a critical evaluation, I conclude that Carter's concerns can be addressed by MEK based on MARVE. Section 6.3.2.4 details the debate between Carter (2019) and Jarvis (2015). Again, the core issue lies with what is to be fixed in a modal assessment. I conclude that MEK can address Carter's and Jarvis's concerns.

I conclude in Section 6.4 that the issues and concerns raised regarding the assimilation of HEC into epistemology can be resolved by MEK based on MARVE.

6.2 Ramifications of HEC for ALVE

This section explores the intersection of HEC in the philosophy of mind and ALVE. The aim is to see whether knowledge can be extended, as in the EM hypothesis, by considering the interactions between an agent and artefacts that are beyond the skull and skin of the agent. To formulate the conditions for EK, Pritchard accepts Clark's formulation of EC.

Pritchard explores potential candidates for extended cognitive process that are analogous to non-extended cognitive process. For the ramifications of ALVE with HEC, Pritchard (2018a) uses the following strategy. As discussed in Chapter 2, extended cognitive processes are consistent with the parity principle. As per Clark and Chalmers (C&C; 1998):

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head,* we would have no hesitation in recognising as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. (C&C, 1998, p. 8)

Pritchard's (2018a) construal of EK, which is essentially the ramifications of ALVE with Clark's EC, can be summarised as:

1. Pritchard accepts Clark's formulation of EC, including the parity principle. This leads to the functional parity of internal and external cognitive processes, as in the functional equivalence of Otto's external memory (his notebook) and Inga's biological memory.

2. Pritchard accepts Clark's glue and trust conditions to avoid cognitive bloat and to give seamless and unreflective cognitive processes. The glue and trust conditions for EC ensure the constitutive role of external artefacts. These conditions are that an artefact is (a) readily accessible, (b) consistently available and (c) automatically endorsed. Pritchard (2018a) argues that the trust and glue conditions ensure that

an external artefact is cognitively integrated with the agent and results in seamless and unreflective cognitive processes, just like the innate cognitive faculties, as normal cognitive processes occur thoughtlessly without significant reflection.

3. Pritchard (2018a) argues that cooperative feedback loops are a feature of cognitive integration.

4. Extended cognitive processes should resemble ordinary unextended cognitive processes. For this, Pritchard proposes cognitive integration, in which extended and unextended cognitive processes should have the same characteristics of fluidity and seamlessness. Ordinary unextended cognitive processes are seamless and fluid, and they do not require reflection by the agent. A cognitive process occurs thoughtlessly in a seamless and fluid way.

5. Once cognitive integration has been achieved, ALVE can accommodate integrated cognitive processes, if they are reliable, i.e. whether they lead to a sufficient degree of cognitive success in the relevant environment. In Pritchard's ALVE, if *S* knows *p*, then *S*'s true belief *p* is the product of a reliable belief-forming process. It is appropriately integrated within *S*'s cognitive character, such that her cognitive success is, to a significant degree, creditable to her cognitive agency (Pritchard, 2010b). According to Pritchard, the general structure of ALVE is as follows: knowledge is a safe belief that arises out of the reliable cognitive traits that make up one's cognitive character, such that one's cognitive success is, to a significant degree, creditable to one's cognitive character.

6. In summary, cognitive integration is crucial for EC to be considered a source of EK because it ensures that external cognitive processes contribute reliably to the formation of true beliefs, while functioning seamlessly with the agent's internal cognitive faculties. The extent and nature of this integration are essential for understanding how EC can contribute to EK, as they ensure that the agent's external cognitive processes are as reliable and effective in knowledge production as their internal counterparts. Cognitive integration is analogous to unextended cognitive processes, such as seamless, fluid, unreflective, thoughtless and functionally equivalent biological cognitive processes.

Pritchard (2018a) argues that for a bona fide case of EK, there should be functional equivalence between seamless EK and the corresponding seamless unextended knowledge. EK arises from extended abilities based on reliable extended cognitive processes that are sufficiently cognitively integrated with the agent, i.e. the level of cognitive integration should support seamless cognition, as in normal cognition.

6.3 Cognitive Integration: EC and EK

Why is cognitive integration important for EK? If an artefact is not properly integrated with the agent, what can go wrong in knowledge acquisition? Consider Lehrer's (1990) counterexample to reliabilism, True Temp.

True Temp: True Temp has (though entirely unbeknownst to him) a temperature-detecting device implanted in his head that regularly produces accurate beliefs about the ambient temperature (extracted from Carter and Kallestrup, 2018, p. 47).

As Carter and Kallestrup conclude:

> As many commentators have accepted, the intuition is strong here that True Temp doesn't attain *knowledge* in the above scenario, even though (thanks to the implanted thermometer) he reliably generates true temperature beliefs, which by the reliabilist's lights is supposed to be all besides truth that matters.
>
> (Carter & Kallestrup, 2018, p. 47)

Here the issue is that the temperature-detecting device is not cognitively integrated with True Temp's cognitive character. The cognitive integration of an external artefact means that the cognitive processes involving the artefact should be synchronised with the cognitive character of the agent. This means that the artefact and the individual's cognitive system should work together in a coordinated and cohesive manner as a form of continuous reciprocal causation (CRC), which is a key principle in the EC framework. CRC refers to the ongoing, dynamic interaction between an individual and an external artefact in which both influence each other in a reciprocal manner. In such a system, the external artefact and the individual's cognitive processes constantly affect and modify each other, leading to an integrated cognitive system that extends beyond the individual's brain. Moreover, seamless fluidity is required in EC, as in normal cognition. The resultant cognitive processes should not be an unusual aspect of the cognitive character of the agent. The true beliefs that True Temp forms are not due to True Temp's ability. However, if True Temp were aware of the temperature-detecting device and if it were cognitively integrated with True Temp's character, then the true beliefs formed by True Temp have to be considered as knowledge.

In the same way, Greco argues that strange and fleeting processes are not part of the cognitive character of an agent:

> For the cognitive faculties and habits of a believer are neither strange nor fleeting. They are not strange because they make up the person's intellectual character – they are part of what make her the person that she is. They are not fleeting because faculties and habits by definition are stable dispositions – they are not the kind of thing a person can adopt on a whim or engage in an irregular fashion. (Greco, 1999, p. 288)

Knowledge comprises true safe beliefs that arise from the ability of an agent such that the exercise of ability is due to the agent's intellectual character. As Breyer and Greco (2008) argue, no strange or fleeting processes can be considered as part of the agent's intellectual character. If there are strange and fleeting processes, a true belief can be reliably formed without justification or knowledge, as in the case of an agent using clairvoyance to form true beliefs reliably, although these do not give rise to justification or knowledge.

Here the question is how can an external artefact become integrated with a cognitive agent? As we have seen in Chapter 2, C&C (1998) claim that an external resource can be considered as a constituent part of cognition only if the resource is reliably available and typically invoked. Any information thus retrieved must be automatically endorsed, and the information within the resource must be easily accessible, as and when required (trust and glue conditions). An external artefact that meets the trust and glue conditions for cognitive processes has a constituent role in the accomplishment of a cognitive task. Therefore, cognition extends beyond the skin bag and skull. The claim that the mind can extend beyond the brain and body requires that external artefacts have a constitutive role in cognition. A causal or enabling role would not be sufficient to establish the extension of the mind into the world. However, even if the constitutive role of external artefacts and the extension of the mind can be established, there is a problem in demarcating the limit of that extension of the mind into the environment via external artefacts. If the words in Otto's notebook have a constitutive role in cognitive processes, then could the numbers in a telephone directory or the results of a Google search also be a part of a cognitive process? If this is the case, then cognition would become rampantly extended via external artefacts and lead to cognitive bloat.

As explained earlier, the objective of Clark's trust and glue conditions is to avoid cognitive bloat, i.e. the rampant expansion of cognitive processes via external artefacts, and to ensure that there is cognitive integration between an external artefact and the agent leading to a kind of unreflective fluidity that is in accord with the characteristics of the agent’s innate cognitive faculties.

Carter and Kallestrup (2020) raise an important question about whether we can count as knowledge the results of whatever cognition is the extracranial epistemic analogue of intracranial cognition, which does lead to knowledge. It is not clear what extent of cognitive integration is required for an external artefact to be part of an agent’s cognitive architecture. The cognitive integration required for extended cognitive processes is referred to as metaphysical cognitive integration, whereas the cognitive integration required for such extended cognitive processes to produce knowledge is referred to as epistemic cognitive integration.

Carter and Kallestrup (2020) argue that the True Temp case satisfies all of Clark’s trust and glue conditions. True Temp’s temperature-detecting device is reliably available and typically invoked, the information retrieved from the device is automatically endorsed, and the information from the temperature-detecting device is easily accessible as and when required. However, the general consensus among epistemologists is that True Temp lacks knowledge, as the device is not properly integrated with True Temp’s cognitive character and does not give True Temp an ability to produce a true belief about the temperature. Although True Temp’s temperature-detecting device does satisfy all of Clark’s trust and glue conditions, these conditions are not sufficient for the level of cognitive integration required for True Temp’s true belief to count as knowledge. Thus, Carter and Kallestrup (2020) propose a “univocal view” for the relationship between metaphysical cognitive integration and epistemic cognitive integration, such that an artefact is metaphysically integrated if and only if it is epistemically integrated.

For cognitive integration and to avoid cognitive bloat, Carter and Kallestrup (2018) add a fourth condition to the three trust and glue conditions provided by Clark: the reliability of the resource must be endorsed (by the agent). Since the reliability of True Temp’s temperature-detecting device is not endorsed by True Temp, the true belief formed by True Temp cannot be considered as knowledge. However, if True Temp endorses the reliability of the temperature-detecting device, then the true belief formed by True Temp would count as knowledge.

However, Aizawa (2018) raises a criticism that the fourth condition fails to solve cognitive bloat, as with a telephone directory. When using a telephone directory, an agent can meet all three trust and glue conditions, as the telephone directory is readily available and typically invoked, the information retrieved is automatically endorsed, and the information within the telephone directory is easily available. The fourth condition is also satisfied, as the agent can endorse the reliability of the telephone directory. If the agent meets all the conditions, cognition can still extend to the entire telephone directory, which is counter-intuitive.

For EK, Palermos accepts the parity principle and the trust and glue conditions. However, Palermos (2011) realises that the trust and glue conditions are not sufficient to solve the coupling-constitution fallacy and cognitive bloat. As per Palermos (2011), in addition to Clark's trust and glue conditions, for an external artefact to have a constitutive role in the overall cognitive mechanism of the agent, CRC between the outer and inner parts are required. Palermos (2011) argues that these 3 +1 criteria ensure the cognitive integration of an external artefact with the agent's overall cognitive mechanism, and thereby, they avoid the coupling-constitution fallacy and cognitive bloat.

Aizawa (2018) argues that even when combined, Clark's three trust and glue conditions, the fourth condition and the CRC proposed by Palermos (2014b) still fail to address Adam and Aizawa's (A&A's) criticisms against Clark's EC, such as the coupling-constitution fallacy and cognitive bloat. A&A argue that EM theorists are making a coupling-constitution error when they suggest that the causal or dependent role of an external artefact is a constitutional role. If an external artefact Y exerts a causal influence on a cognitive process X, that does not mean that Y is part of X. A&A (2010) assert that even an acceptance of the coupling of external features with cognition does not mean that cognition extends to every part of that system. A non-biological external artefact, such as a pen or paper, cannot be considered a constituent part of cognitive processes. A&A (2010) raise the problem of cognitive bloat, i.e. if we accept the extension of the mind into the environment, then what is the extent of that extension? Can any external object contribute to cognition? A&A argue that if an external artefact can be considered as contributing to cognition because it has a constitutive role, then many external artefacts could too, resulting in cognitive bloat. Clark's reply to these criticisms was based on the parity principle and the functional parity of the external artefacts, as in the functional parity of Otto's notebook and Inga's biological memory. Rupert (2004) argues that his theory of embedded cognition can easily accommodate the extended processes in a causal enabling role rather than a constitutive role.

I argue in Chapter 2 that Clark’s formulation of EC, which is based on the parity principle, the trust and glue conditions, and functional parity, fails to address the criticisms raised by A&A and Rupert regarding the enabling versus the constitutive role of external artefacts. Unfortunately, Pritchard’s formulation of EK is based on Clark’s formulation of EC. Aizawa (2018) rightly indicates that criticisms such as the coupling-constitution fallacy and cognitive bloat are applicable to HEC and EK.

Pritchard’s acceptance of Clark’s formulation of EM, with functional parity, the trust and glue conditions, and its claim that EC is analogous to unextended cognition, makes Pritchard’s EK prone to the criticism relating to the enabling versus the constitutive role of external artefacts.

To overcome A&A’s criticisms, namely the coupling-constitution fallacy and cognitive bloat, I argue in Chapter 3 that to establish the constitutive role of an external artefact in cognition, EC should be based on dynamical systems theory, i.e. there is a requirement for a feedback loop between the agent and the artefact and for the manipulation of that artefact by the agent. I agree with Palermos that the CRC of the inner and outer parts is required to establish the constitutive role of an external artefact. Although I disagree with Palermos’s EC example of an agent and TVSS, I agree with Palermos that to establish the constitutive role of an external artefact, there is a potential application of dynamical systems theory that is a candidate for EC in which the agent manipulates the external artefact to achieve cognitive success via a feedback loop. I argue in Chapter 3 that a closely coupled non-linear relation between an external artefact and a cogniser cannot alone be used to establish the constitutive role of an external artefact in cognition in a principled way. The necessary condition for EC is that the resultant feedback loop between the agent and the external artefact should result in the manipulation of external information-bearing structures when accomplishing a cognitive task. Such a closely coupled non-linear feedback loop could establish the constitutive role of an external artefact in cognition.

My conclusive argument in Chapter 3 is that functionalism based on parity is neither necessary nor sufficient for EM. However, a necessary and sufficient condition for EM is the back-and-forth dynamic interaction of the agent with an external artefact and the processing of the information gained from that dynamic interaction in realising a cognitive task. I agree with Palermos (2014a) that “the only requirement for an external element to count as a constitutive part of the agent’s cognitive system is that it be non-linearly related to the rest of the agent’s cognitive system” (Palermos, 2014, p. 10).

6.3.1 MEK Based on Modified EM

As mentioned earlier, Pritchard’s construal of EK based on ARVE has problems with respect to the enabling versus constitutive role of external artefacts and cognitive bloat. As discussed in Chapter 2, the criticisms raised by A&A and Rupert regarding the coupling-constitution fallacy and cognitive bloat are applicable to Pritchard’s EK, as it is based on the parity principle and the trust and glue conditions for the constitutive role of an artefact and for cognitive integration between an artefact and an agent. Pritchard’s core conditions for an EC candidate to become EK is that the EC candidate must be sufficiently integrated with the agent’s cognitive character so that the extended cognitive abilities can form true beliefs in an unreflective way. The conditions for EK can easily be met by the enabling role of an external artefact.

Moreover, Pritchard does not clearly describe the relationship between an artefact and an agent. As explained in Chapter 4, the modal account of risk (MAR) cannot accommodate the full spectrum of risks associated with knowledge production. The risk associated with an artefact further complicates any risk evaluation in knowledge production. In summary, to address the criticism against ARVE and to make ARVE the preferred account of knowledge, the following gaps need to be addressed:

1. MAR is not adequate, as it requires an account of risk that can cover the entire spectrum of risk in knowledge production, especially when an artefact is involved.

2. The cognitive integration between an external artefact and an agent must be established by considering the role of the artefact and its relationship with the agent during knowledge production. The parity principle and the trust and glue conditions are unable to address the coupling-constitution fallacy and cognitive bloat.

My focus here is to establish MEK such that it is immune to the criticism of the enabling versus constitutive role of external artefacts while maintaining the epistemic externalist criteria of the seamless cognitive integration of external artefacts with an agent’s cognitive character, which results in extended cognitive abilities. Modified extended cognition, which is discussed in Chapter 3, can be a candidate for EK and it can be accommodated within MARVE, as discussed in Chapter 5.

Here, I rely on the dynamic relationship between an artefact and an epistemic agent, as discussed in Chapter 3. I argued in Chapter 3 that the existence of dynamic interactions, such as feedback loops with external information-bearing structures and the manipulation of those external information-bearing structures, is a sufficient and necessary condition for the constitutive role of an external artefact in cognition. If the cogniser and external artefact both have an active role in cognition via a feedback loop, I argue that the overall system comprising the cogniser, the external information-bearing structure and the feedback loop are the constituent parts of cognition. This role of an external information-bearing structure via a feedback loop cannot be considered to be an enabling role. If we consider only the enabling role of the external artefact in cognition and undermine the feedback loop with the cogniser, then we lose the dynamic interaction of the cogniser in the manipulation of the external artefact and the reciprocal causation of the manipulated external information-bearing structure on the cogniser. Further, we lose the changes in the cognitive processes that occur over time due to the interactions between the external information-bearing structure and the cogniser via the feedback loop. Furthermore, in such cases, the cognitive task cannot be accomplished without the feedback loop. Such a construal of modified extended cognition is not based on the parity principle or the trust and glue conditions. Now, the question is how can such a modified candidate for EC become EK? To realise EK, the cognitive integration of the external artefact with the agent is required if the agent is to form true beliefs unreflectively. How is such cognitive integration possible with a dynamic construal of EC?

As I discuss in Section 3.2.1 of Chapter 3, the relationship between an artefact and a cognitive agent is based on various types of loops, such as open loops, feedback loops, feedforward loops and emulation loops. I argue in Chapter 3 that such a dynamic relationship between an artefact and an agent can result in the transformation of the loops over time. Once a feedback loop has become established, as in modified extended cognition, and integrated with the agent, it can be transformed into a cognitive ability to form true beliefs in a seamless, fluid, unreflective way, just like the exercise of cognitive faculties. I argue in Chapter 3 that the feedback loops involving external artefacts and a cognitive agent can diffuse the problem of the enabling versus constitutive role of external artefacts in producing EK, if the external artefacts are sufficiently integrated with the agent’s cognitive character such that this results in extended cognitive abilities that form true beliefs in an unreflective, thoughtless way.

In summary, Pritchard's use of the parity principle and the trust and glue conditions is unable to address the coupling-constitution fallacy and cognitive bloat. Pritchard's ARVE has limitations in terms of capturing the risk due to an external artefact being involved in knowledge production. MAR, as discussed in Chapter 5, which includes both the modal and probabilistic components of risk, can cover the entire spectrum of risk in knowledge production. In MARVE, the role of an artefact and its relationship with an agent are based on the dynamic relationship between them, i.e. various loops. Thereby, this avoids any reliance on the parity principle and the trust and glue conditions for cognitive integration. MARVE, as discussed in Chapter 5, can accommodate EK and is immune to the coupling-constitution fallacy and cognitive bloat.

In the new version of EK (MEK), only two conditions are required for EK: (1) a feedback loop and (2) seamless cognitive integration. For example, knowledge gained using night vision goggles (Pritchard, 2018a), as described in Chapter 3, is a candidate for EK. When a soldier uses night vision goggles for the first time, the relationship between the agent and the device is like that between a subject and an instrument. When the subject uses the instrument, a feedback loop is set up (e.g. what she sees will guide where she looks, so she learns to adjust the settings of the device to suit her preferences, and so on). However, over time, the soldier will become familiar with the instrument and will completely integrate it with her overall cognitive processes in a seamless fashion so that she forms beliefs unreflectively. The seamless cognitive processes involving night vision goggles are a candidate source of EK. This example shows that a candidate source of EK via a highly reflective route, where the initial feedback loops help to enhance the soldier's cognitive ability, is such that, over time, the googles can become integrated with the soldier's overall cognitive processes and the soldier can then achieve cognitive success unreflectively. There is no bio-prejudice in this approach. It does not segregate mind and body or perception and action. Everything is integrated.

Compared with Pritchard's construal of EK, MEK has the following advantages:

- MEK is immune to criticisms like the coupling-constitution fallacy and cognitive bloat as it is based on a dynamic relationship between an artefact and an agent rather than the parity principle and the trust and glue conditions.

- A risk evaluation in the hybrid account of risk uses both the probabilistic and modal components for the risk assessment. The hybrid account of risk can capture the full spectrum of risk in knowledge production, including the risks associated with artefacts. Pritchard’s MAR cannot account for the risks associated with artefacts (see Chapter 5 for more details).

The following section considers various discussions as well as issues with and potential solutions for EK based on MEK and MARVE, which was described in Chapter 5. MARVE can encompass MEK. For extended cognitive processes, there has to be cognitive integration via a feedback loop between the external artefact and the agent and the manipulation of that external artefact by the agent. As per MARVE, the factors necessary for the production of knowledge are as follows: (1) agential factors, (2) factors relating to epistemic tools and (3) environmental factors. As discussed in Chapter 5, the MARVE framework explains what is to be fixed in a modal assessment: (1) agential factors and (2) the performance of an epistemic tool and its dynamic relation with the agent. What can vary in a modal assessment is the environment.

6.3.2 Debates and Issues with EK

This section focuses on various debates around EK, such as the debates between Kelp and Vaesen and those between Carter and Jarvis.

6.3.2.1 EK: Debate between Vaesen and Kelp

Vaesen (2011) argues that there are cases in which an agent’s positive epistemic dependence due to environmental factors can achieve cognitive success, which can count as knowledge. In such cases, robust virtue epistemology (RVE) fails because the cognitive success is not primarily attributed to the cognitive abilities of the agent. Vaesen argues that such mundane extended cognitive cases in which the credit for part of an agent’s cognitive success can be attributed to an extended cognitive aid are inconsistent with creditability virtue epistemology (CVE) because CVE attributes the cognitive success solely to the agent’s cognitive ability. Vaesen (2011) argues that CVE fails in mundane extended cognitive cases, such as the example of the baggage inspector Sissi, whose cognitive success in detecting a bomb was not primarily credited to Sissi’s cognitive ability since part of the credit can be attributed to Sissi’s manager, who had installed a system producing false alarms to make the baggage inspector more vigilant.

Kelp’s (2013b) initial response claims that the problem with CVE is a familiar old problem, irrespective of whether an internal or external cognitive aid takes part in the cognitive success of the agent. Thus, he refutes Vaesen’s argument that the agent’s cognitive success due to positive epistemic dependence via EC is incompatible with CVE.

To do so, Kelp (2013) provides a counterexample involving Sissi*, whose vigilance has been enhanced by a vigilance-increasing drug developed by her manager. In this case, Sissi*'s cognitive success is partly credited to her internal cognitive ability, which has been augmented by medicine rather than EC. In such cases, Kelp argues that the agent's cognitive success is not solely credited to the agent's cognitive ability, rather it can partly be attributed to the medicine that enhanced the agent's cognitive ability.

Vaesen (2013) counter-argues that Kelp misses the dilemma that Sissi's cognitive success is not solely credited to Sissi's cognitive ability. According to Vaesen, for cognitive success, CVE requires that Sissi's cognitive success must be primarily creditable to Sissi's exercise of cognitive ability. However, in this case, success is not solely credited to Sissi's cognitive ability. Therefore, CVE fails. If Kelp does not accept that dilemma, Kelp must explain how CVE can account for the success.

Kelp (2014) argues that such environmental aspects in the exercise of cognitive ability can be captured by Sosa's interpretation of virtue epistemology (VE). Kelp (2014) responds to Vaesen, noting that in a normal environment, Sosa's interpretation of VE is consistent with the agent's cognitive success being due to the manifestation of the cognitive ability of the agent. Sosa's interpretation of cognitive ability has three components: (1) inner (IN), (2) competence (CO) and (3) situational (SI). To achieve cognitive success, all three components are required together. This is termed aptness. The SI part ensures that the exercise of cognitive ability has occurred in the right environment, which thereby avoids the pitfalls of any variations in the environment and the subsequent cognitive failure. For Sissi, Kelp argues that the SI aspect is satisfied with the new scanner, and with her IN (increased concentration, etc.), Sissi forms true belief that the suitcase contains a bomb; therefore, Sissi's cognitive success may manifest competence.

The core of the debate is about whether RVE can be an adequate theory of knowledge for the cases where the agent's cognitive success is not primarily due to the cognitive ability of the agent, such as Sissi. Only the strong interpretation of RVE can address Gettier cases, i.e. cognitive success is primarily credited to the cognitive ability of the agent. However, for Sissi, such a strong attribution of credit is not possible because Sissi's cognitive success is partly due to the new bag scanner. Here, I do not go into the details of whether RVE is adequate. Rather, my intention is to assess the relevance of EC and EK for Sissi. In this debate, I would like to highlight one of the aspects that demonstrate that Sissi's case is not a candidate for EC. The new bag scanner is a cognitive aid. This artefact has an enabling role and does not meet the requirement for a constitutive role in Sissi's cognitive processes. Sissi's knowledge cannot be considered as a candidate for EK, as the knowledge does not meet the criteria for EK that I elaborated earlier, namely (1) a feedback loop and (2) cognitive integration. Sissi's case is an example of an enabling or causal role of a cognitive aid in the agent's cognitive processes. In a footnote, Vaesen mentions that he is unsure whether the example can be considered as producing EK. Kelp's initial attempt was to show that the problem raised by Vaesen was not specific to EC, as Vaesen claims. However, in a subsequent response, Kelp (2014) accepts the dilemma in Sissi's case and explains that Sosa's interpretation of VE is immune to such cases. I agree with Kelp that Sosa's interpretation of VE is immune to the criticism that Sissi's cognitive success is not primarily credited to the exercise of her cognitive ability. Rather, part of the credit goes to the new scanner. As per Sosa's VE, Sissi manifests competence with the new scanner (SI) and applies her IN.

6.3.2.2 Kelp's Version of Extended Epistemology

Kelp (2014) formulates a counterexample in which the existing VE theories and the modal account of knowledge both fail. In this example, a *time seeker* looks at a stopped clock and forms beliefs about the time. However, a *timekeeper* knows that the clock is stopped, so he ascertains the actual time using two reliable clocks and confirms that the time shown by the stopped clock is currently the correct time. According to Kelp, the timekeeper is a reliable informant, since if the time on the stopped clock was wrong, the timekeeper would have corrected the time seeker's belief about the time. In this case, the time seeker forms a true belief about the time. Kelp argues that this scenario cannot be captured by RVE. For example, if there were no timekeeper, as in Sosa's interpretation of VE, the time seeker would have false belief of the time, as the SI factor of his interpretation of VE cannot be met. The SI factor ensures that under normal conditions, i.e. if the clock were working properly, the time seeker would have formed a true belief about the time. However, with the presence of the timekeeper, i.e. a reliable informant, the time seeker forms true beliefs irrespective of the SI factor, as it remains the same for the clock. Kelp, thus, concludes that this interpretation of VE fails in such scenarios.

Kelp considers a modally close world in which the time seeker looks at the clock before or after the timekeeper looks at it. In this case, whatever her competence in reading the clock, the time seeker would form a false belief of the time because she cannot form knowledge about the time by looking a stopped clock. However, if the time seeker and timekeeper look at the clock together, i.e. when the time seeker looks at the stopped clock, the timekeeper verifies that the time shown is the actual time, then the time seeker would form a true belief about the time due to her clock-reading competence, since if the time shown was incorrect, the timekeeper would inform the time seeker. Kelp concludes that the modal account of safety fails in such a scenario, as the clock-reading competence of the agent stays the same, but the agent forms true or false beliefs depending on whether the timekeeper is present.

Kelp argues that: "There are cases, notably *The Timekeeper*, that pose a problem for even our most promising traditional accounts of knowledge, but can be dealt with nicely by (at least a certain type of) extended epistemology" (Kelp, 2014, p. 247).

Kelp (2014) proposes a new version of EK to resolve the time seeker/timekeeper case. Kelp explores whether the time seeker/timekeeper can be considered as involving EC. Kelp (2014) argues that:

> It is the time seeker's way of belief-formation or method that extends beyond her skin. The time seeker's method includes an on-board part and an external part. While the on-board part is a standard clock-reading process, the external part, which is contributed by the timekeeper, is a monitoring process that would alert the time seeker to inaccuracies of clock readings.
>
> (Kelp, 2014, p. 244)

Kelp concludes that this case is an example of EC. It satisfies the parity principle because the time seeker's clock-reading competence is an on-board part that is analogous to the timekeeper's monitoring process, which is an external part of the time seeker's overall cognitive processes. Moreover, Kelp considers that the time seeker and timekeeper are reliably coupled. Kelp notices that there is a potential for decoupling and that it may be problematic to consider such cases as extended. However, Kelp relies on Clark and Wilson's proposal for transient extended cognitive systems, which allows the time seeker/timekeeper case to be considered as EC because there is a transient extended cognitive system in which the extended systems involve temporary or transient forms of cognitive augmentation. With such systems, the timekeeper in a modally close world is a reliable informant for the time, irrespective of when the time seeker looks at the stopped clock. Thus, the time seeker forms a true belief about the time. Kelp claims that his formulation of extended epistemology yields the right result, namely that the time seeker knows the time. I think that Kelp's formulation of EC is not aligned with C&C's EC. The standard formulation of EC requires that an external artefact has a constitutive role in the overall cognitive processes of the agent. In this case, however, there is no such artefact, and the attribution of knowledge by the extended cognitive processes is based on the testimony of the timekeeper, who is a reliable informant about the time.

The discussion so far regarding EC considers a cognitive agent with an external artefact. However, the example of a time seeker and a timekeeper is not aligned with the standard formulation of EC. It is more aligned with the domain of testimony.

Since the *relevant* initial conditions and belief-forming processes are fixed in all modally close worlds, the time seeker, the timekeeper and the stopped clock all function the same as they do in the actual world. That is, the time seeker will form a true belief about the time due to the positive epistemic dependence on the reliable informant (the timekeeper), irrespective of when the time seeker looks at the stopped clock. If the time seeker comes later or earlier than the time shown on the clock, the timekeeper will warn her that the clock is stopped and tell her the right time. If a reliable informant like the timekeeper is part of the agent's initial belief-forming process, then in a modally close world, the timekeeper still has a role in the belief-forming process of the time seeker. Therefore, the cases where the timekeeper comes later or earlier than the time seeker do not comply with the modal requirement of fixed initial conditions.

Kelp could raise a possible objection to this scenario, as Kelp would say that it is very unclear what counts as relevant. Kelp would argue that we can avoid all counterexamples if "relevant" picks out only worlds in which the belief is true. What would be a non-trivial way of unpacking what is relevant? Kelp could also point out this aspect for Barney. If in the fake barn case we hold fixed that Barney is looking at a real barn, then his belief will be safe. Now, we might say that the condition that he is looking at a real barn is not one of the relevant initial conditions. But the question arises why not? Note that the answer cannot be that Barney clearly does not know, as that would render the account circular. One way of proceeding is by taking the method of belief formation as one of the relevant initial conditions. That will get the fake barn's case right, but it would not be enough for the timekeeper, as Kelp argues.

As we have seen in Chapter 5, the production of EK involves: (1) agential factors, (2) factors relating to epistemic tools and (3) environmental factors. Agential factors include cognitive faculties such as memory, reasoning, introspection, perception, cognitive traits, cognitive character and disposition. Chapter 5 details what relevant initial conditions are to be fixed for a sensible safety assessment, namely agential factors are to be fixed and the performance of an epistemic tool and its dynamic relation with the agent are to be fixed. What can vary in a close possible word is the environment. Therefore, if the timekeeper, who is a reliable informant, is present when the time seeker forms a true belief about the time, then this is relevant in the actual world and this relevant initial condition must be fixed in a modal assessment. If, instead, we consider that the timekeeper is part of the environment and whether they are present can vary in close possible worlds, then the true belief formed by the time seeker in the actual world is epistemically fragile, as the environment is allowed to change in a modal assessment. However, in my understanding, Kelp considers that the timekeeper is not part of the environment. Rather, this is a scenario with a reliable informant, as in testimonial cases.

As per MEK, Kelp's claim that the time seeker's belief-forming process extends beyond her skin bag cannot be supported. External artefacts, such as a stopped clock and a reliable informant, have a causal or enabling role in the agent's overall cognitive processes. In MEK, only two conditions are required for EK: (1) a feedback loop and (2) unreflective cognition.

MEK is immune to the criticism relating to the enabling versus constitutive role of external artefacts. Feedback loops involving external artefacts and a cognitive agent can diffuse this problem in producing EK if the external artefacts are sufficiently integrated with the agent's cognitive character such that this results in extended cognitive abilities that form true beliefs in an unreflective, thoughtless way. To establish the role of feedback loops in producing EK, I have explained the nature of feedback loops based on (1) dynamical systems theory and (2) the cognitive integration of EC with an agent's cognitive character in unreflective cognition.

In conclusion, Kelp's construal of extended epistemology is, in fact, not a candidate for producing EK. The stopped clock and timekeeper aide the cognition of the time seeker rather than having a constitutive role in the time seeker's cognitive processes.

6.3.2.3 Issues with Environmental Veritic Luck and C&C's Metaphysical Parity

In this section, I expound Carter's argument regarding the potential issues with environmental epistemic luck when C&C's radically extended cognition (REC) is assimilated into epistemology.

Carter (2013b) argues that there are concerns with mainstream epistemologies when they attempt to assimilate REC. As an extension of the parity principle to epistemology, Carter proposes epistemic parity (E-parity). However, Carter identified an obvious tension between metaphysical parity (M-parity) and E-parity. Carter raises a fundamental issue with C&C's REC for EK. As I discuss earlier, the fundamental tenet of C&C's EC is the parity principle, which is a basic requirement for the functional equivalence of the cognitive processes that occur in the head and those that occur when an external artefact takes part in the overall cognitive processes of the agent. Carter (2013b) argues that C&C's parity prevents metaphysical bio-prejudice but not epistemic bio-prejudice, which he called M-parity. In line with C&C's parity principle, to prevent epistemic bio-prejudice, Carter (2013b) formulated E-parity: "For agent *S* and belief *p*, if *S* comes to believe *p* by a process which, were it to go on in the head, we would have no hesitation in ascribing knowledge of *p* to *S*, then *S* knows *p*" (Carter, 2013b, p. 4204). M-parity means that extracranial cognitive processes are analogous to intracranial cognitive processes. E-parity means that extracranial epistemic processes are analogous of whatever intracranial forms of cognition we count as forming knowledge.

Carter (2013b) formulates a counterexample for Otto with two cases, which are like the fake barn case:

Case 1: Otto has Alzheimer's. A jokester changed all the entries in Otto's diary, making them an hour earlier, except for Otto's doctor's appointment, which was left unchanged. When Otto subsequently looked at his diary, he formed a true belief about his doctor's appointment.

Case 2: Otto** does not have Alzheimer's. One day, he forgot about all his appointments, except for his doctor's appointment. With his clear biological memory, he remembered his doctor's appointment.

Carter (2013b) argues that case 1 is like the case of a fake barn, because when Otto looks at his diary, he can form a true belief about his doctor's appointment but his cognitive success cannot be considered as knowledge because in a modally close world, Otto may look at other entries in the diary and form false beliefs about his other appointments. Carter claims that if we agree that Otto's cognitive success regarding his doctor's appointment cannot be counted as knowledge, then M-parity is in trouble, because it requires that extracranial cognitive processes are analogous to intracranial cognitive processes. Therefore, M-parity cannot ensure the safety of knowledge derived from extracranial cognitive processes. On the other hand, if we grant that Otto's cognitive success regarding his doctor's appointment is knowledge, then this contradicts the ALVE conditions for the safety of knowledge.

Carter (2013b) further argues that cases 1 and 2 are equivalent in terms of E-parity, in which extracranial knowledge is analogous to intracranial knowledge. Therefore, if case 1 cannot be considered as knowledge, then case 2 also cannot be considered as knowledge. However, Carter argues that in mainstream epistemology, case 2 is an example of knowledge, as it is derived from Otto**'s clear biological memory. These examples illustrate the apparent tension between M-parity and E-parity.

Carter (2013b) concludes that: "The no knowledge verdict in cases where an agent's correctness is down to environmental luck is at odds with both Clark and Chalmers' original parity principle as well as with the E-parity." As a solution to this impasse, Carter proposes re-examining what is to remain fixed in a modal assessment. Carter also suggests a re-examination of the hypothesis in EC that environmental luck can undermine the safety of a cognitive success gained by extended cognitive processes.

I agree with Carter that there is an obvious tension between M-parity and E-parity. M-parity cannot ensure the safety of knowledge production, as in case 1, where Otto's cognitive processes are extended as per M-parity. However, this knowledge cannot be considered as a candidate for EK because in a modally close world, Otto* may form a false belief about the doctor's appointment: "In many close possible worlds, the jokester *does not* overlook the doctor's appointment entry, and Otto is an hour early to the appointment" (Carter, 2013b, p. 7).

As we discuss in Chapter 5, the following factors need to be considered if EC is to be a candidate source for EK: (1) agential factors, (2) factors relating to epistemic tools and (3) environmental factors. The relevant initial conditions to be fixed in a modal assessment include agential factors and the performance of the epistemic tool and the dynamic relation between the epistemic tool and the agent. What can vary in close possible words is the environment. The jokester in case 1 is like the potential issue with the environment in fake barn country, as the role of the jokester can be varied in close possible worlds. This makes case 1 epistemically fragile, i.e. the epistemic risks are high because in a modally close world, the jokester changes the doctor's appointment and Otto forms a false belief about his doctor's appointment. Case 1 is analogous to the barn façade case. In both cases, knowledge is not gained because of the high epistemic risk.

The issue with E-parity for cases 1 and 2 is related to the different cognitive processes associated with the cases. In case 1, Otto must look at his diary to remember his doctor's appointment, but in case 2, Otto** remembered his doctor's appointment using his biological memory. In a modal assessment, the initial conditions must be fixed along with the belief-forming processes. For case 1, perception and the subsequent remembering are both belief-forming processes, whereas in case 2, only memory is required and there is no perceptual input. For case 1, the relevant initial belief-forming processes, such as perception and remembering, are to be fixed and what can vary is the environment, i.e. the jokester can change the doctor's appointment in close possible worlds. For case 1, Otto could look at a different entry in his diary in close possible worlds and then form a false belief. Therefore, case 1 cannot produce knowledge. However, for case 2, the belief-forming process is remembering, and it is to be fixed in close possible worlds. Since memory can produce true beliefs that count as knowledge, then in nearby close possible worlds, Otto without Alzheimer's can form a true belief about his doctor's appointment. Therefore, case 2 can produce knowledge. Based on this, I disagree with Carter that case 1 is an extracranial analogue of case 2, as the cognitive processes in cases 1 and 2 are different. Therefore, Carter's concern about M-parity and E- parity dissolves. I share the concerns raised by Carter that M-parity does not ensure the safety of the target belief, which is required for EK. However, the question is whether parity is required for EC or not. As I argued earlier, parity is not required for EC. As discussed, MEK is not based on the parity principle. Instead, it is based on the dynamic relationship between an agent and an artefact. In MEK, only two conditions are required for EK: (1) a feedback loop and (2) unreflective cognition. The above example of knowledge gained using night vision goggles is a candidate for EK. This example has a highly reflective route, as the initial feedback loop helps to enhance the soldier's cognitive ability. There is no bio-prejudice in this approach. It does not segregate mind and body or perception and action. Everything is integrated. Since M-parity is not required for EK, E-parity is also not required, so that there is no obvious tension in assimilating EC in epistemology.

In summary, parity is not required for either modified EM or MEK. A potential candidate for EC can be checked to see if there are feedback loops and the manipulation of an external artefact. Such a case of EC can be a potential candidate for EK, if the external artefact is properly integrated with the cognitive character of the agent. Therefore, Carter's argument about the potential tension between M-parity and E-parity is not relevant for MEK, as it is based on feedback loops, the manipulation of external artefacts and cognitive integration.

It is possible that the conditions for EC, such as a feedback loop and unreflective cognition, will also not ensure the safety of the target belief, as required for EK. As discussed in Chapter 5 regarding MARVE, when an external artefact is involved in knowledge acquisition, the safety of the target belief has to be checked case by case. It is rather difficult to establish a set of conditions that ensure the safety of any target belief formed by EC.

6.3.2.4 Debate between Jarvis and Carter: Jarvis's Argument against Carter

In this section, I expound on the debate between Carter and Jarvis regarding the potential issues with environmental epistemic luck when C&C's REC is used with mainstream epistemology.

Jarvis (2015) argues that there are no concerns in accommodating REC in mainstream epistemologies. He states that the guiding idea behind REC is that the cognitive roles undertaken by internal entities could be done by radically extended entities. Conversely, Jarvis (2015) argues that if this is correct, the cognitive roles done by radically extended entities could be done instead by internal biological entities. Therefore, Jarvis's strategy is to produce counterexamples to show how radically extended cognitive roles can be replaced by internal biological entities.

Jarvis (2015) modifies Carter's case 2 (i.e. Otto** without Alzheimer's) so that it is like case 1. Unbeknownst to Otto**, a bio-jokester uses pharmaceuticals and subliminal suggestions to create false memories about Otto**'s appointments that day such that he believes that all his appointments that day are one hour earlier than they really are, except that the jokester overlooked Otto**'s doctor's appointment and left it unchanged. Now both case 1 (i.e. the jokester-altered Otto with Alzheimer's) and case 2 (the bio-jokester-altered Otto** without Alzheimer's) are identical with respect to environmental luck.

Jarvis (2015) argues that epistemological theories should not have special carve-out clauses for REC. Since an application of traditional epistemological theories yields the same results for cases 1 and 2, then there is no requirement for special clauses to deal with REC. In other words, assimilating REC is not an issue for traditional epistemologies, as Carter worries. Jarvis claims that: "Any epistemological theory that is materially adequate when restricted to cases without REC will be materially adequate *tout court*."

Carter (2019) disagrees with Jarvis, arguing that the environmental luck (veritic epistemic luck) differs between the jokester case and the bio-jokester case. Case 1 (the jokester case) is like the barn façade case. It is a clear case of knowledge-undermining epistemic luck. However, for the bio-jokester case, Otto**'s knowledge regarding his doctor's appointment is safe. Therefore, Jarvis's conclusion that cases 1 and 2 are epistemically symmetrical is incorrect. If one were to adjust the bio-jokester case, then it would not be consistent with either intervening or environmental epistemic luck. However, the jokester case is a clear example of environmental epistemic luck. Carter (2019), thus, concludes that Jarvis's argument fails and the issues with assimilating EC into mainstream epistemology persist.

As discussed earlier, the relevant initial conditions to be fixed in a modal assessment are agential factors and the performance of the epistemic tool and the dynamic relation of the epistemic tool with the agent. What can vary in close possible words is the environment. In Chapter 5, we discuss aetiological functions, which are the normal functions of an item in normal conditions. Aetiological functions must be fixed as initial conditions of the belief-forming process to allow a meaningful modal assessment of risk. In the bio-jokester case, the aetiological functions are not fixed as the biological configuration of Otto** has been modified with a drug. We cannot make a meaningful modal assessment if we change the biological configuration of the cognitive agent. Thus, Jarvis's (2015) bio-jokester fails to comply with what is to be fixed for a modal assessment. What can be changed in a modal assessment is the interface between the agent and the environment, as explained in Chapter 5.

I disagree with Carter that the issues with assimilating EC into epistemology persist, i.e. the tension between M- parity and E-parity persists. Since my proposal for MEK is not based on the parity principle, the concern about M-parity and E- parity dissolves.

I share the concern raised by Carter (2013b) that for EC to be a candidate for EK:

> It is hard to maintain intuitive judgments about safety, and furthermore, it becomes less clear how we should distinguish between environmental and intervening epistemic luck, a distinction that is much more straightforward in

> cases where what is fixed under the description of a cognitive process is the intracranial cognitive process employed in the actual world.
>
> (Carter, 2013b, p. 326)

As mentioned earlier, it is possible that the conditions for EC, such as a feedback loop and unreflective cognition, will also not ensure the safety of the target belief, as required for EK. As discussed in Chapter 5, regarding MARVE, when an external artefact is involved in knowledge acquisition, the safety of the target belief has to be checked case by case. It is rather difficult to establish a set of conditions that ensure the safety of any target belief formed by EC.

6.4 Conclusions

In summary, Pritchard's MAR has limitations as it cannot capture the full spectrum of risk in knowledge production, especially when an artefact is involved, as the risks associated with an artefact in knowledge production require both the probabilistic and modal components. Pritchard's ALVE has limitations in terms of establishing and differentiating the relationship and interface between an artefact and an agent and the relationship and interface between the environment and an agent. Pritchard claims that ALVE is a preferred account of knowledge in assimilating EC for EK. However, its reliance on the parity principle and the trust and glue conditions makes ALVE prone to the criticisms raised by Adams, Izawa and Rupert regarding the coupling-constitution fallacy and cognitive bloat. As Carter argues, a reliance on the parity principle has an obvious tension when EC is assimilated into epistemology, as M-parity prevents bio-prejudice, but the equivalent E-parity cannot. This necessitates a requirement to modify ARVE, which I did in Chapter 5 by establishing the nature and relationship of the interface between an artefact and an agent and the interface between the environment and an agent. MARVE is derived from modified EC, which is based on feedback loops and the manipulation of external artefacts by an agent. Since MARVE is not based on the parity principle, the concern raised by Carter regarding M-parity and E-parity dissolves. MARVE is immune to the criticisms raised by Adams, Izawa and Rupert, as MARVE can clearly establish the constitutive role of an external artefact via a dynamic feedback loop between an agent and the artefact. MEK, which is based on MARVE and the hybrid account of luck, is immune to the coupling-constitution fallacy and cognitive bloat. The issues raised in various debates about EK, such as those between Vaesen and Kelp and between Carter and Jarvis, can be addressed by MEK based on the hybrid account of risk, MARVE and the dynamic relations among an agent, an artefact and the environment.

7 CONCLUSIONS

As explored in various chapters, this thesis primarily focuses on externalism in the philosophy of mind, specifically the extended mind (EM) thesis, and externalism in epistemology, as exemplified by anti-luck virtue epistemology (ALVE). It also examines the implications of EM and ALVE in producing extended knowledge (EK). The thesis investigates the dynamic interaction between cognitive agents and external artefacts. It proposes a modified form of the EM thesis based on dynamical systems theory (DST) to address the coupling-constitution fallacy and cognitive bloat. The construal of modified EM based on DST has the potential for multidisciplinary integration and leads to an integrated version of EM, which integrates elements of DST, niche construction theory (NCT), cognitive niche theory (CN), cognitive niche construction theory (CNC) and developmental systems theory to provide insights into the historical patterns of external artefacts. This multidisciplinary framework contrasts with the traditional, Cartesian-inspired view of cognitive internalism, in which the mind is seen as distinct and internal to the agent. The assimilation of DST-based EC into epistemology emphasises the importance of external artefacts in knowledge acquisition. The modal account of risk (MAR) seems inadequate for addressing the entire spectrum of risk when knowledge acquisition involves artefacts. Therefore, I recommend a hybrid account of risk (HAR) that has modal and probabilistic components to cover the entire spectrum of risk in knowledge acquisition involving external artefacts. By using DST-based extended cognition (EC) and HAR, I proposed modified anti-risk virtue epistemology (MARVE) and modified EK (MEK).

In Chapter 2, I discussed EC as proposed by Clark and Chalmers (C&C; 1998) to address potential criticisms. The traditional view, rooted in Cartesian thought, posits a distinct and entirely internal mind. C&C (1998), however, propose a different perspective in which, during the formation of beliefs, the mind can extend into the environment through closely coupled external artefacts. They present the EM hypothesis, which is exemplified by the role of a pen and paper in complex computations. The chapter explores the parity principle put forth by C&C, which suggests that if a part of the world functions as a process that would be recognised as cognitive if it occurred in the head, then it should be considered part of the cognitive process.

Adams and Aizawa (A&A; 2010) argue against this, claiming that C&C are incorrect in asserting that external items like a pen and paper can play a constitutive role in cognitive processes. They refer to this error as the coupling-constitution fallacy and maintain that the causal dependence of a cognitive process on an external artefact does not imply that the artefact is a part of the cognition process. A&A assert that even if there is a coupling of external features with cognition, that does not mean that cognition extends to every part of that system.

Similarly, Rupert (2009) expresses concerns about the constitutive role of external artefacts. Although he acknowledges the enabling role of external artefacts in complex computations, Rupert argues that cognitive processes are wholly internal to the agent, being biological and occurring in the brain and body. Consequently, no external artefact can be considered a part of cognition. The criticisms by A&A and Rupert challenge EM theorists by indicating that the proponents of the EM thesis, as formulated by C&C, lack a principled method to establish the constitutive role of an external artefact. This challenge includes the coupling-constitution fallacy and cognitive bloat. Although C&C (1998) define cognitive processes through the parity principle, critics like Rupert and A&A do not accept this definition, as they consider that cognitive processes to be entirely biological and internal to the agent.

I conclude that with Clark's formulation of EM, it is difficult to address these criticisms. Proponents of EC, such as Rowlands, Menary and Sutton, oppose the common-sense functionalism employed by C&C. However, as explained in Chapter 2, these alternative versions of EC still struggle to address the coupling-constitution fallacy and cognitive bloat. Considering the potential of DST, a framework for EC based on DST could potentially overcome issues like the coupling-constitution fallacy and cognitive bloat without necessitating the controversial stance of functionalism. This is the focus of Chapter 3.

Modified extended mind: Chapter 3 primarily focuses on the necessary and sufficient conditions for EC based on DST. It is supported by various arguments from developmental systems theory and NCT and an analysis of patterns and historicism in external information-bearing structures along with the role of pattern recognition. The chapter explores the integration of EC and NCT to explain how intellectual abilities arise from the innate cognitive abilities of humans, as endowed by evolution.

The DST-based EM approach effectively addresses the coupling-constitution fallacy and cognitive bloat. The EC thesis, when based on DST, offers a general framework for theorising about dynamic systems. It elucidates how the relationship between an artefact and an agent can constitute a coupled system. According to DST, for an external artefact to play a constitutive (rather than merely auxiliary) role in cognition, there must be feedback loops between the cogniser and the artefact, with the cogniser actively manipulating the artefact. In such feedback loops, the perception and action of the agent become inseparable. Pritchard’s (2018a) example of a soldier using night vision goggles (NVGs) illustrates this point well: the perception and action of the soldier in this case are closely coupled and inseparable. This relationship between the artefact and the agent, which is modelled on a feedback loop in DST, forms the foundation of the EC thesis. Although both the NVGs and tactile–visual sensory substitution (TVSS) scenarios provide insights into EC, the NVGs scenario is a more direct and active exploration of how external artefacts can become deeply integrated into our cognitive processes. The immediacy of the consequences, the active manipulation and the enhancement of a natural sense endow the NVGs scenario with more explanatory potential. By emphasising the dual conditions of manipulation and feedback, one can argue for a richer, more dynamic and deeper interactive understanding of EC that captures the complexity of human–tool interactions more comprehensively than relying on feedback loops alone. Combining manipulation with feedback offers a more robust framework for understanding the intricate interplay between agents and their artefacts.

The modified EM addresses the coupling-constitution fallacy, which occurs when there is confusion between an external artefact being merely coupled (connected or associated) with a cognitive process and being a constitutive part of that process. Critics argue that just because a cognitive process is coupled with an external artefact, it does not necessarily mean the artefact is part of the cognitive process itself. The modified EM is immune to this fallacy because it specifies a more rigorous criterion: the existence of a feedback loop involving active manipulation. This goes beyond mere coupling by ensuring that the artefact is not just associated with the cognitive process but is actively involved and essential to it. The requirement for active manipulation within a feedback loop in the modified EM clearly delineates when an external artefact transitions from being a mere tool (coupling) to an integral component of cognition (constitution). A feedback loop implies that there is a dynamic, reciprocal interaction between the agent and the artefact, in which both the cogniser and the artefact are mutually influential. This level of integration is specific to the context of the feedback loop, which is a targeted, task-specific interaction in which the artefact becomes a constitutive part of the cognitive process for the duration and purpose of that specific task.

Cognitive bloat occurs when the boundaries of the mind are over-expanded because too many external elements are considered as being part of cognitive processes. Standard EM frameworks could lead to the problematic implication that almost any external object an individual interacts with could be considered part of their cognition. The modified EM avoids cognitive bloat by setting a clear criterion for what counts as a constitutive part of cognition: the presence of a feedback loop involving the manipulation of an external information-bearing structure. This criterion is specific and restrictive, thus preventing the indiscriminate inclusion of external artefacts in cognitive processes. Only artefacts that are actively manipulated in a feedback loop, thereby having a direct and significant impact on the cognitive task, are considered as being able to extend cognition. The nature of the interaction in the feedback loop is what grants the artefact its constitutive role.

Although Rupert's argument for simplicity in favour of the hypothesis of embedded cognition is compelling, that approach could oversimplify complex cognitive phenomena. In cognitive science, the most parsimonious explanation is not always the most accurate or comprehensive, especially when dealing with complex, dynamic systems. The modified EC provides a more nuanced understanding of the interplay between cognisers and their environments by recognising the deep integration of external artefacts in cognitive processes. This comprehensive approach may be necessary to fully explain the intricacies of how cognition extends beyond the brain.

Intersection of EC, NCT, CN, CNC and virtue epistemology (VE) and the formulation of an integrated version of EM: In Chapter 3, I argue that dynamic interactions, such as feedback loops and the manipulation of external information-bearing structures, play a crucial role in how external artefacts contribute to cognition. Although I theorise that these interactions are both necessary and sufficient for the constitutive role of external artefacts in cognition, it is important to consider the broader context provided by recent developments in various fields. The application of DST to cognition, coupled with insights from evolutionary biology through NCT, NC, CNC and developmental systems theory, aligns with the principles of the EC hypothesis. These fields emphasise the importance of both internal and external factors in shaping cognitive processes and suggest that human cognition is not solely an internal, brain-based process but also extends into the environment so that it is shaped by how humans modify their surroundings.

Since a dynamic feedback loop between an agent and an external artefact is a common theme in both NCT and DST-based EC, there is the potential to integrate new developments in evolutionary biology, such as NCT, into EC. Recognising that external artefacts can have a constitutional role, as opposed to merely an enabling role, is significant in addressing the limitations of cognitive internalism. This traditional view posits that cognitive processes occur solely within an individual's mind and struggles to account for the complexity of human cognition, especially in social interactions and cultural developments. Social cognition often involves a reliance on and interactions with external artefacts, which play a fundamental role in shaping cognitive processes in complex tasks and social contexts. The integration of NCT, EC, CN and CNC leads to an interdisciplinary framework that enhances our understanding of human thinking and provides insights into the evolution and history of artefacts and their relationships with agents. NCT, with its focus on ecological inheritance, complements the emphasis of EC on the feedback loops between an agent and artefacts in cognition. This integration offers a comprehensive framework for understanding cultural inheritance and the dynamic relationships between artefacts and agents.

The ongoing interaction emphasised in NCT, in which organisms actively modify their environments, aligns with the principles of EC. This concept is further enriched by VE, particularly its focus on intellectual virtues as sophisticated cognitive traits developed through interactions with the environment. NCT acknowledges the significant role that individual organisms play in shaping their environment, which impacts evolutionary pressures and contributes to ecological changes. The intersection of NCT, EC and VE provides a holistic understanding of cognition, which highlights the importance of external interactions and cultural influences in cognitive development. This approach underlines the dynamic and adaptive nature of human cognition, which is shaped by both internal capacities and external environmental interactions.

The integration of NCT and EC offers a powerful framework for explaining the development and history of artefacts and the relationships between artefacts and agents. This integrated approach increases the explanatory power compared to classifying artefacts based on functional or intentional roles alone and provides deeper insights into the classification, history and development of artefacts. In conclusion, this integrated approach with NCT and EC not only enhances our understanding of the cognitive development of humans and the evolution of artefacts but also provides a more nuanced understanding of the nature of relationships between cognitive agents and artefacts. It represents a significant step forward in explaining the complexity of human cognition and its interplay with the environment and cultural development.

In Chapter 4, I discuss post-Gettier epistemology. I analyse knowledge acquisition and explore how it can be enriched by the modified EM approach. Traditional knowledge analysis, such as the justified true belief (JTB) model, follows a tripartite structure. However, Gettier-style cases demonstrate how lucky true beliefs can undermine the concept of knowledge. This chapter primarily focuses on Gettier cases, various accounts of luck, anti-luck epistemology (ALE), RVE, ALVE and criticisms of ALVE. The nature and characteristics of luck and its relationship to knowledge remain contentious topics in epistemology. Various theories of luck, including the probabilistic, lack of control and modal accounts, are discussed. However, these theories have counterexamples, and none provides a fully adequate explanation of luck. Yet, there is consensus that knowledge is incompatible with certain types of luck.

Pritchard formulates ALVE by selecting an account of luck that ensures the reliability of target beliefs and a virtue-theoretic condition to demonstrate that cognitive success is attributable to the cognitive abilities of the agent. Pritchard employs a modal account of luck (MAL) to ensure the safety of target beliefs, such that acquired knowledge is not due to luck. Many epistemologists, including Lackey, Carter, Peterson and De Grefte, have criticised MAL, arguing that it does not encompass the full spectrum of luck in knowledge acquisition. Recently, Pritchard shifted from focusing on luck to emphasising risk. He argues that risk is a fundamental and forward-looking concept, whereas luck is backward-looking, assessing what went wrong. Consequently, Pritchard modified MAL to MAR and accordingly revised his theory of knowledge from ALVE to anti-risk virtue epistemology (ARVE).

Novel account of risk for EK: In Chapter 5, my goal is to enrich risk assessments in ARVE, particularly when external artefacts are involved in knowledge acquisition. I argue that MAR alone is insufficient to capture the full spectrum of risks associated with external artefacts that are involved in knowledge acquisition. I propose that a hybrid account of risk (HAR), combining both probabilistic and modal aspects of risk, is necessary to comprehensively capture these risks. This hybrid account effectively encompasses the full spectrum of risks associated with knowledge acquisition involving external artefacts. To enrich the risk assessment, I evaluate the dynamic relationships among a cognitive agent, an artefact and the environment and propose a comprehensive framework for risk assessments in ARVE.

When incorporating an external artefact into knowledge acquisition, the assessment of risk involves evaluating both the functionality of the artefact and its impact on knowledge attribution. Risks include the possibility that the artefact will malfunction, which can produce false beliefs, and the potential for the artefact to contribute to the formation of false beliefs even when functioning correctly. This comprehensive risk assessment considers both the technical reliability of the artefact and the contextual factors influencing belief formation. In engineering, risk assessments traditionally focus on the likelihood of failure and the severity of its consequences. However, this approach can sometimes overlook the complex and nuanced vulnerabilities of a system.

HAR, which integrates modal intuition with probabilistic methods, addresses the complexities of real-world scenarios more effectively. It encompasses both the probability of an event occurring and the severity of its impact to provide a more nuanced understanding of risk, especially in knowledge acquisition involving external artefacts. This approach is particularly relevant for complex systems with inherent uncertainties. The modal and probabilistic components of HAR address the full spectrum of risk associated with knowledge acquisition, making it a suitable model for risk assessments in both epistemology and engineering. In applying HAR in engineering, it is crucial to consider both the number and significance of changes from normal operating conditions. This approach enhances the reliability of safety assessments, especially in scenarios characterised by high uncertainties.

Based on the modified EC and the new HAR, I have modified Pritchard's ARVE to create MARVE and proposed a modified version of EK.

MARVE and EK: As we have seen, Pritchard's MAR has limitations in that it cannot capture the full spectrum of risk in knowledge production, particularly when an artefact is involved. An assessment of the risks associated with an artefact in knowledge production requires both probabilistic and modal components. HAR, which encompasses both modal and probabilistic elements, can address this shortfall. Pritchard's ARVE also falls short in establishing and differentiating the relationships and interfaces between an artefact and an agent and between the environment and an agent. I address this gap by establishing the nature of these interfaces in MARVE, which integrates feedback loops and artefact manipulation from EC, as these are crucial for understanding how artefacts contribute to cognitive processes and knowledge acquisition. This integration enables MARVE to evaluate the risks associated with artefacts in a more nuanced manner that reflects their role in extended cognitive systems.

It is important to differentiate between agent–tool and agent–environment interfaces in a risk assessment. This differentiation is crucial for understanding how cognitive processes interact with external elements. If an artefact results in EC, then the artefact must be cognitively integrated with the agent and impact cognitive processes in a specific, functional manner. In contrast, the environment encompasses a broader range of external factors that may influence cognitive outcomes. This distinction is important for accurately evaluating the likelihood and impact of cognitive failures in a modal risk assessment. The distinction between the agent–tool and the agent–environment interfaces is crucial, not for undermining the concept of cognitive extension but for recognising the distinct ways in which tools and environmental factors can interact with cognitive processes. Making an accurate risk assessment of cognitive tasks, particularly when external artefacts are involved, necessitates understanding these interactions.

HAR, with its probabilistic and modal components, can capture the full spectrum of risks, making it applicable to both epistemology and engineering. This is particularly important when an artefact is involved in knowledge acquisition. In engineering, risk assessments typically focus solely on probability, but the inclusion of a modal component can capture broader risks, especially in the presence of uncertainty.

HAR is especially crucial in the context of EK due to the intricate nature of the interaction between an agent and external artefacts that are integral to the process of knowledge acquisition. In EC scenarios, these external artefacts are not simply tools utilised by an agent; rather, they become integral parts of the cognitive process itself. EK scenarios encompass both modal risks (pertaining to various possible scenarios) and probabilistic risks (related to the likelihood or chance of events). It is possible that these artefacts will fail (probabilistic risk) or that they will not operate as expected across different scenarios (modal risk).

In contrast, in scenarios involving enabling artefacts, such as those found in non-EK acquisition, the role of the artefact is more straightforward. Consequently, the modal aspects of risk, which consider how different potential scenarios could impact knowledge acquisition, are typically less complex.

How MARVE fares better than Pritchard's ARVE: Traditional epistemological approaches do not adequately address the potential risks introduced by external artefacts in knowledge acquisition. However, as discussed, external artefacts can play a crucial role in extending cognitive processes into the environment. In Chapter 5, I discuss extended cognitive processes as potential sources of EK, such that the external artefact plays a constitutive role. Pritchard's ARVE lacks the tools to assess the risks due to such artefacts. MARVE, however, can assess these risks more effectively.

For example, consider the soldier using NVGs, as explored in Chapter 2. I argue that this scenario, modelled using Beer's (2000) framework, represents EC because cognition and behaviour are inseparable in such an integrated system with feedback loops. Pritchard considers this a suitable candidate for EK. However, his modal account does not address the risks associated with the NVGs themselves. MARVE, through HAR, can assess these risks more comprehensively. For instance, the reliability of brand-new goggles bought from a reputable manufacturer may have a low failure rate, thereby ensuring the safety of beliefs formed using the goggles. In contrast, older and less reliable goggles present a higher potential for false beliefs, which highlights the need for a more comprehensive risk assessment.

In summary, MARVE provides a more refined framework for risk assessments in knowledge acquisition, especially in cases involving EC. It addresses the limitations of traditional epistemology by incorporating a more comprehensive understanding of the risks associated with external artefacts. The integration of NVGs into the soldier's cognitive process, for example, demonstrates the need for a holistic approach to understanding how external artefacts integrate with and influence cognitive processes. HAR's dual focus on probability and modal closeness offers a comprehensive framework for assessing risks, considering both the likelihood of failure and the impact of slight changes in circumstances. This approach aligns with the EC framework and acknowledges the active role of external artefacts in cognitive processes and their impact on epistemology.

Modified extended knowledge: The conclusion of the thesis emphasises that none of the existing theories of luck fully captures the spectrum of luck in knowledge acquisition. Pritchard's MAR fails to address the full spectrum of risk, especially when artefacts are involved. My modification of ARVE (MARVE) addresses these limitations by establishing a clear relationship between artefacts, agents and the environment. MARVE, derived from the principles of the modified EM and not constrained by the parity principle, overcomes the criticisms of Adams, Aizawa and Rupert. It establishes the constitutive role of an external artefact via a dynamic feedback loop.

MEK, which is based on MARVE, HAR and modified EM, is immune to criticisms such as the coupling-constitution fallacy and cognitive bloat. It provides a comprehensive framework for addressing the issues raised in various debates about EK and relies on the effective integration of HAR with the dynamic relations among agents, artefacts and the environment.

REFERENCES

Abraham R. H. & Shaw, C. D. (1992). *Dynamics – The Geometry of Behavior, 2nd edn*. Addison-Wesley.

Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14(1), 43–64.

Adams, F. & Aizawa, K. (2005). Defending non-derived content. *Philosophical Psychology*, 18, 661–9.

Adams, F. & Aizawa, K. (2010). Defending the bounds of cognition. In R. Menary (Ed.), *The Extended Mind*. MIT Press, Cambridge, MA.

Aizawa, K, (2018). Extended cognition, trust and glue, and knowledge. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos & D. Pritchard (Eds.), *Extended Epistemology, 1st edn* (Chapter 3). Oxford University Press, Oxford.

Bach-y-Rita, P. & Kercel, S. W. (2003). Sensory substitution and the human–machine interface. *Trends in Cognitive Science*, 7(12), 541–6.

Basalla, G. (1988). *The Evolution of Technology*. Cambridge University Press.

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72, 173–215.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Science*, 4, 91–9.

Beer, R. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11, 209–43.

Begley, S. (2001). How it all starts inside your brain. *Newsweek*, 137(7), 40.

Bertolotti, T., Bertolotti, T., Magnani, L. & Magnani, L. (2017). Theoretical considerations on cognitive niche construction, *Synthese*, 194(12), 4757–79.

Boyd, R., Richerson, P. J. & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2), 10918–25.

Braddon-Mitchell, D. & Jackson, F. (2007). *The Philosophy of Mind and Cognition, 2nd edn*, Blackwell Publishing, Oxford; Malden, MA.

Breyer, D. & Greco, J. (2008). Cognitive Integration and the Ownership of Belief: Response to Bernecker, *Philosophy and Phenomenological Research*, 76(1), 173–84.

Burge, T. (1979). Individualism and the mental. In French, P. et al. (Eds.), *Midwest Studies in Philosophy, Volume 4, Metaphysics*. University of Minnesota Press, Minneapolis.

Carter, J. A. (2013a). A problem for Pritchard's robust virtue epistemology. *Erkenntnis*, 78(2), 253–75.

Carter, J. A. (2013b), Extended cognition and epistemic luck. *Synthese*, 190(18), 4201–14.

Carter, J. A. (2016). Robust virtue epistemology as anti-luck epistemology: A new solution. *Pacific Philosophical Quarterly*, 97(2016), 140–55.

Carter, J. A. (2019). Epistemic luck and the extended mind. In I. M. Church & R. J. Hartman (Eds.), *The Routledge Handbook of the Philosophy and Psychology of Luck* (pp. 318–30), Routledge.

Carter, J. A. & Kallestrup, J. (2018). Extended circularity: a new puzzle for extended cognition. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos & D. Pritchard (Eds.), *Extended Epistemology, 1st edn*. Oxford University Press, Oxford.

Carter, J. A. & Kallestrup, J. (2020). Varieties of cognitive integration. *Noûs (Bloomington, Indiana)* 54(4), 867–90.

Carter, J. A. & Peterson, M. (2017). The modal account of luck revisited. *Synthese* 194, 2175–84.

Carter, J. A., Kallestrup, J., Palermos, S. O. & Pritchard, D. (2014). Varieties of externalism. *Philosophical Issues*, 24(1), 63–109.

Carter, J. A., Palermos, S. O. & Gordon, E. C. (2016). Extended emotion. *Philosophical Psychology*, 29(2), 198–217. DOI: 10.1080/09515089.2015.1063596

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O. & Pritchard, D. (2018). *Extended Epistemology, 1st edn*, Oxford University Press, Oxford.

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

Chemero, A. & Silberstein, M. (2008). Defending extended cognition. In V. Sloutsky, B. Love & K. McRae (Eds.). *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 129–34). Psychology Press.

Chisholm, R. M. (1977). *Theory of Knowledge, 2nd edn*. Prentice-Hall, Englewood Cliffs, NJ.

Churchland, P. S. (2002). *Brain-wise*. MIT Press, Cambridge, MA.

Clark, A. (1997a). *Being There: Putting Brain, Body, and World Together Again*. MIT Press Cambridge, MA.

Clark, A. (1997b). The Dynamical Challenge. *Cognitive Science*, 241, 461–81.

Clark, A. (1998). Time and mind. *Journal of Philosophy*, XCV (7), 354–76.

Clark, A. (1999). Minds, brains and tools (with a response by Daniel Dennett). In H. Clapin (Ed.), *Philosophy of Mental Representation* (pp. 66–90). Clarendon Press.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language*, 16(2), 121–45.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, Oxford.

Clark, A. (2005). Word, niche and super-niche: How language makes minds matter more. *Theoria. An International Journal for Theory, History and Foundations of Science*, 20(3), 255–68.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8), 370–4.

Clark, A. (2008a). Pressing the flesh: A tension in the study of the embodied, embedded mind? *Philosophy and Phenomenological Research*, *76* (1), 37 – 59.

Clark, A. (2008b). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, Oxford.

Clark, A. (2010). Coupling, constitution, and the cognitive kind. In R. Menary (Ed.), *The Extended Mind* (pp. 81–101). MIT Press, Cambridge, MA.

Clark, A. & Chalmers, D. (1998). The extended mind. In R. Menary (Ed.), *The Extended Mind* (pp. 27–43). MIT Press, Cambridge, MA.

Clark, A. & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, 7(1), 5–16.

Coughanowr, R. D. (1991). *Process Systems and Analysis Control*. McGraw-Hill International Editions.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience.* Clarendon Press, New York, Oxford University Press, Oxford.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. G. P. Putnam's Sons.

Dawkins, R. (2004). Extended phenotype – but not too extended. A reply to Laland, Turner and Jablonka. *Biology and Philosophy,* 19(3), 377–96.

De Grefte, J. (2019). Pritchard versus Pritchard on luck. *Metaphilosophy* 50(1–2), 3–15.

De Grefte, J. (2020). Towards a hybrid account of luck. *Pacific Philosophical Quarterly*, 101(2), 240–55.

Dennett, D. (1987). *Intentional Stance.* MIT Press, Cambridge, MA.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88 (1), 27–51.

Dennett, D. C. (1996). *Kinds of Minds*. Basic Books.

Eliasmith, C. (2001). Attractive and in-discrete: A critique of two putative virtues of the dynamicist theory of mind. *Minds and Machines*, 11, 417–26.

Farina, M. & Lavazza, A. (2022). Incorporation, transparency and cognitive extension: Why the distinction between embedded and extended might be more important to ethics than to metaphysics. *Philosophy & Technology,* 35(1), 10.

Favela, L. H., Amon, M.J., Lobo, L. & Chemero, A. (2021). Empirical evidence for extended cognitive systems. *Cognitive Science*, 45(11), e13060-n/a.

Flynn, E. G., Laland, K. N., Kendal, R. L. & Kendal, J. R. (2013). Target article with commentaries: Developmental niche construction. *Developmental Science*, 16(2), 296–313.

Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research*, 25–26, 4–12.

Gelder, T. van (1995). What might cognition be if not computation? *Journal of Philosophy*, 92(7), 345–81.

Gelder, T. van (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615–28.

Gettier, E. (1963). Is justified true belief knowledge? In *Analysis* (pp. 121–3).

Gibson, J. G. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston.

Giere, R. N. & Moffatt, B. (2003). Distributed cognition: Where the cognitive and the social merge. *Social Studies of Science*, 33(2), 301–10.

Ginet, C. (1975). *Knowledge, Perception and Memory*. D. Reidel Publishing Company, Dordrecht.

Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy* 73(20), 771– 91.

Graham, P. J. (2012). Epistemic entitlement. *Noûs*, 46(3), 449–82.

Greco, J. (1999). Agent reliabilism. *Philosophical Perspectives*, 13, 273–96

Greco, J. (2003). Virtue and luck, epistemic and otherwise. *Metaphilosophy,* 34(3), 353–66.

Greco, J. (2007). Worries about Pritchard's safety. *Synthese*, 158(3), 299–302.

Greco, J. (2008). What's wrong with contextualism? *The Philosophical Quarterly,* 58(232), 416–36.

Greco, J. (2009). Knowledge and success from ability. *Philosophical Studies,* 142(1), 17–26.

Greco, J. (2010). *Achieving Knowledge.* Cambridge University Press, Cambridge.

Greco, J. (2012a). Recent work on testimonial knowledge. *American Philosophical Quarterly*, 49(1), 15–28.

Greco, J. (2012b). A (different) virtue epistemology. *Philosophy and Phenomenological Research*, 85(1), 1–26.

Greco, J. (2020). *The Transmission of Knowledge*. Cambridge University Press, Cambridge.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *The Behavioral and Brain Sciences*, 27(3), 377–96.

Gupta, M., Prasad, N., Dey, G., Joshi, S. & Vidya, A. (2017). Niche construction in evolutionary theory: The construction of an academic niche? *Journal of Genetics*, 96(3), 491–504.

HSE (2001). *Reducing Risks, Protecting People: HSE's Decision-making Process*. HSE Books, Sudbury.

Hurley, S. L. (1998). *Consciousness in Action*. Harvard University Press.

Hurley, S. (2010). The varieties of externalism. In R. Menary (Ed.), *The Extended Mind*. MIT Press, Cambridge, MA.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press, Cambridge, MA.

IEC 61508 (2010). Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems, Parts 1–7, IEC 61508, 2nd Edn. Geneva, Switzerland: International Electrotechnical Commission.

IEC 61511 (2016). Functional Safety: Safety Instrumented Systems for the Process Industry Sector, Parts 1-3, IEC 61511, Geneva, Switzerland: International Electrotechnical Commission, 2016.

Izquierdo E. J. & Beer R. D. (2016). The whole worm: Brain–body–environment models of *C. elegans*. *Current Opinion in Neurobiology*, 40, 23–30.

Jablonka, E. & Lamb, M. (2005). *Evolution in Four Dimensions*. MIT Press, Cambridge, MA.

Jarvis, B. (2015). Epistemology and radically extended cognition. *Episteme*, 12(4), 459–78.

Kalivas, P. W. & Volkow, N D. (2005). The neural basis of addiction: A pathology of motivation and choice. *American Journal of Psychiatry*, 162(8), 1403–13.

Kallestrup, J. & Pritchard, D. (2014). Virtue epistemology and epistemic twin earth. *European Journal of Philosophy*, 22(3), 335–57

Kallestrup, J. & Pritchard, D. (2016). From epistemic anti-individualism to intellectual humility. *Res Philosophica*, 93(3), 533–52

Kelp, C. (2009). Pritchard on virtue epistemology. *International Journal of Philosophical Studies*, 17(4), 583–7.

Kelp, C. (2011). In defence of virtue epistemology. *Synthese*, 179(3), 409–33.

Kelp, C. (2012). Anti-luck virtue epistemology. *Grazer Philosophische Studien*, 2013, 211–25.

Kelp, C. (2013a). How to motivate anti-luck virtue epistemology. *Grazer Philosophische Studien*, 88(1), 211–25.

Kelp, C. (2013b). Extended cognition and robust virtue epistemology. *Erkenntnis* 78(2), 245–52.

Kelp, C. (2013c). Knowledge: The safe-apt view, *Australasian Journal of Philosophy*, 91(2), 265–78.

Kelp, C. (2014). Extended cognition and robust virtue epistemology: Response to Vaesen. *Erkenntnis* 79(3), 729-732.

Kelso, S. (1995). *Dynamic Patterns: The Self-organization of Brain and Behavior*. MIT Press, Cambridge, MA.

Kerr, B. (2007). Niche construction and cognitive evolution. *Biological Theory*, 2(3), 250–62

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4), 513–49.

Lackey, J. (2008). What luck is not. *Australasian Journal of Philosophy*, 86(2), 255–67.

Lackey, J. (2009). Knowledge and credit. *Philosophical Studies*, 142(1), 27–42.

Laland, K. N. & O'Brien, M. J. (2011). Cultural niche construction: an introduction. *Biological Theory*, 6(3), 191–202

Laland, K. & Sterelny, K. (2006). Perspective: Seven reasons (not) to neglect niche construction. *Evolution*, 60(9), 1751–62.

Laland, K., Matthews, B. & Feldman, M. (2016). An introduction to niche construction theory. *Evolutionary Ecology*, 30(2), 191–202.

Lehrer, K. (1990). *Theory of knowledge.* Routledge, London.

Levin, J. (2008). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (fall 2008 edition), http://plato.stanford.edu/archives/fall2008/entries/ functionalism/.

Levins, R. & Lewontin, R. (1985). *The Dialectical Biologist*. Harvard University Press.

Lewontin R. C. (1983). Gene, organism, and environment. In D. S. Bendall (Ed.), *Evolution from Molecules to Men* (pp. 273–285). Cambridge University Press, Cambridge, UK.

Lewontin R. C. (2000). *The Triple Helix: Gene, Organism, and Environment*. Harvard University Press, Cambridge, USA.

Menary, R. (2010). Cognitive integration and extended mind. In R. Menary (Ed.), *The Extended Mind* (pp. 227–45). MIT Press, Cambridge, MA.

Morris, A. S. & Langari, R. (2012). *Measurement and Instrumentation: Theory and Application.* Academic Press.

Nussbaum, M. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

Odling-Smee, J. & Laland, K. N. (2011). Ecological inheritance and cultural inheritance: What are they and how do they differ? *Biological Theory*, 6(3), 220–30.

Odling-Smee, J. & Turner, J. (2011). Niche construction theory and human architecture. *Biological Theory,* 6(3), 283–9.

Odling-Smee J., Laland, K. & Feldman, M. (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton University Press.

Otis, J. M., Namboodiri, V. M. K., Matan, A. M., Voets, E. S., Mohorn, E. P., Kosyk, O., McHenry, J. A., Robinson, J. E., Resendez, S. L., Rossi, M. A. & Stuber, G. D. (2017). Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. *Nature*, 543(7643), 103–7.

Palermos, S. O. (2011). Belief-forming processes, extended. *Review of Philosophy and Psychology* 2(4), 741–65.

Palermos, S. O. (2014a). Loops, constitution, and cognitive extension. *Cognitive Systems Research*, 27, 25–41.

Palermos, S. O. (2014b). Knowledge and cognitive integration. *Synthese (Dordrecht)* 191(8), 1931–51.

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, 107(Suppl 2), 8993–8999.

Pinker, S. (2014). *Language, Cognition, and Human Nature* (Chapter 13). Oxford University Press.

Pritchard, D. (2005). *Epistemic Luck*, Oxford University Press.

Pritchard, D. (2006). Greco on reliabilism and epistemic luck. *Philosophical Studies*, 130(1), 35–45.

Pritchard, D. (2007). Anti-luck epistemology. *Synthese*, 158, 277–97.

Pritchard, D. (2010a). Cognitive ability and the extended cognition thesis, *Synthese*, 175(Suppl 1), 133–51.

Pritchard, D. (2010b). Anti-luck virtue epistemology. In A. Haddock, A. Millar and D. Pritchard (Eds), *The Nature and Value of Knowledge: Three Investigations.* Oxford University Press.

Pritchard, D. (2012). Anti-luck virtue epistemology. *Journal of Philosophy*, 109, 247–79.

Pritchard, D. (2014). The modal account of luck. *Meta Philosophy*, 45(4), 594–619.

Pritchard, D. (2015a). Anti-luck epistemology and the Gettier problem. *Philosophical Studies*, 172(1), 93–111

Pritchard, D. (2015b). Risk. *Metaphilosophy*, 46(3), 436–61.

Pritchard, D. (2015c), Epistemic dependence, *Philosophical Perspectives*, 29(1), 305–24.

Pritchard, D. (2016a). *Epistemology, 2nd edn*. Palgrave Macmillan, London.

Pritchard, D. (2016b). Epistemic risk. *The Journal of Philosophy*, 113(11), 550–71.

Pritchard, D. (2018a). Extended epistemology. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos & D. Pritchard (Eds), *Extended Epistemology, 1st edn* (Chapter 5). Oxford University Press, Oxford.

Pritchard, D. (2018b). Extended virtue epistemology. *Inquiry*, 61(5–6), 632–47.

Pritchard, D. (2018c). Anti-luck virtue epistemology and epistemic defeat. *Synthese*, 195(7), 3065–77.

Pritchard, D. (2020). Anti-risk epistemology and negative epistemic dependence. *Synthese*, 197(7), 2879–94.

Pritchard, D. (2021). Anti-luck epistemology and pragmatic encroachment. *Synthese*, 199(1), 715–29.

Putnam, H. (1975). The meaning of meaning. In K. Gunderson (Ed.), *Language, Mind and Knowledge*. University of Minnesota Press, Minneapolis.

Rescher, N. (2014) The machinations of luck. *Metaphilosophy*, 45(4/5), 620–6.

Riggs, W. D (2009). Luck, knowledge, and control. In A. Haddock, A. Millar, A. & D. Pritchard, D., *Epistemic Value* (Chapter 9). Oxford University Press, Oxford.

Riggs, W. D. (2014). Luck, knowledge, and “mere” coincidence. *Metaphilosophy*, 45(4–5), 627–39.

Ross, D. & Ladyman, J. (2010). The alleged coupling-constitution fallacy and the mature sciences. In R. Menary (Ed.). *The Extended Mind* (pp. 189–226). MIT Press, Cambridge, MA.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, 22(1), 1–19.

Rowlands, M. (2010a). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press, Cambridge, MA.

Rowlands, M. (2010b). Consciousness, Broadly Construed. In Menary, R (Ed.), *The Extended Mind*, (pp. 271–94). MIT Press, Cambridge, MA.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101(8), 389–428.

Rupert, R. (2009). Representation in extended cognitive systems: Does the scaffolding of language extend the mind? In R. Menary (Ed.), *The Extended Mind*. MIT Press, Cambridge, MA.

Rupert, R. (2010). *Cognitive Systems and the Extended Mind*. Oxford University Press, New York.

Schilling, M., Hoinville, T., Schmitz J. & Cruse, H. (2013). Walknet, a bio-inspired controller for hexapod walking. *Biological Cybernetics*, 107, 397–419.

Shapiro, L. (2009). Review of Fred Adams and Ken Aizawa. The bounds of cognition. *Phenomenology and the Cognitive Sciences*, 8, 267–73.

Shapiro, L. (2012). Embodied cognition. In *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press.

Shapiro, L. (2013). Dynamics and Cognition. *Minds and Machines*, 23, 353–75.

Simion, M. (2016). Perception, History and Benefit. *Episteme*, vol. 13, no. 1, pp. 61-76.

Solomon, R. C. (2003). *Not Passion's Slave: Emotions and Choice*. Oxford University Press.

Sosa, E. (2007). *A Virtue Epistemology*. Clarendon Press, Oxford.

Sosa, E. (2009). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*, Vol. 1. Oxford University Press, Oxford.

Sosa, E. (2015). *Judgment and Agency*. Oxford University Press, Oxford.

Spivey, M. (2007). The continuity of mind. New York: Oxford University Press.

Sprevak, M (2009). Extended cognition and functionalism. *Journal of Philosophy*, 106(9), 503–27.

Stapelberg, R. F. (2009). *Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design*. Springer, London.

Sterelny, K. (2001). Niche construction, developmental systems, and the extended replicator. In S. Oyama, P. E. Griffiths & R. D. Gray (Eds.), *Cycles of Contingency: Developmental Systems and Evolution*. MIT Press, Cambridge, MA.

Sterelny, K. (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Blackwell.

Sterelny, K. (2004). Externalism, epistemic artefacts and the extended mind. In R. Schantz (Ed.), *Current Issues in Theoretical Philosophy, Vol. 2: The Externalist Challenge*. Walter de Gruyter.

Sterelny, K. (2006). The evolution and evolvability of culture. *Mind & Language*, 21(2), 137–65. https://doi.org/10.1111/j.0268-1064.2006. 00309.x

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press, Cambridge, MA.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press, Cambridge.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.) *The Extended Mind* (pp. 189–226). MIT Press, Cambridge, MA.

Tedeschi, F. & Carbone, G. (2014). Design issues for hexapod walking robots. *Robotics*, 3(2), 181–206.

Thelen, E. (1995). Motor development: A new synthesis. *American Psychologist*, 50, 79–95.

Thelen, E. & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA.

Thelen, E., & Smith, L. B. (2006). Dynamic systems theories. In W. Damon & R. M. Lerner (Eds.), *Handbook of Child Psychology: Theoretical Models of Human Development* (Chapter 6, pp. 258–311). John Wiley & Sons Inc.

Tooby, J. & DeVore, I. (1987). The reconstruction of hominid evolution through strategic modeling. In W. G. Kinzey (Ed.), *The Evolution of Human Behavior: Primate Models*. SUNY Press.

Tribble, E. (2005). Distributing cognition in the globe. *Shakespeare Quarterly, 56*, 135 – 155.

Vaesen, K. (2011). Knowledge without credit, exhibit 4: Extended cognition. *Synthese*, 181(3), 515–29.

Vaesen, K. (2013). Critical discussion: Virtue epistemology and extended cognition: A reply to Kelp and Greco. *Erkenntnis*, 78(4), 963–70.

Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, 92(7), 345–81.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615–28.

Van Gelder, T. & Port, E. (1995). It’s about time: An overview of the dynamical approach to cognition. In R. Port & T. van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition* (pp. 1–44). MIT Press, Cambridge.

von Mises, R. V. (1957). *Probability, Statistics, and Truth*. Dover Publications, Inc.

Vygotsky, L. (1930). The instrumental method in psychology. In R. van der Veer (Ed.), *The Collected Works of L. S. Vygotsky. Volume 3: Problems of the Theory and History of Psychology*. Plenum Press.

Vygotsky, L. (1934). *Thought and Languag*e. (A. Kozulin, Ed. and Trans.; revised edition). MIT Press, Cambridge, MA.

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press, Cambridge, MA.

Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), *The Extended Mind* (pp. 245–70). MIT Press, Cambridge, MA.

Wheeler, M. & Clark, A. (2008). Culture, embodiment and genes: Unravelling the triple helix. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1509), 3563–75.

Willey, R. J. (2014). Layer of protection analysis. *Procedia Engineering*, 84, 12–22.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9, 625–36.

Wilson, R. A. (2005). Collective memory, group minds, and the extended mind thesis. *Cognitive Process*, 6, 227–36.

Wilson, R. & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition, Cambridge Handbooks in Psychology* (pp. 55–77). Cambridge University Press, Cambridge.