Epistemic Opacity: A Feature Not a Bug

An exploration into the relationship between brains and ANNs

By Keldt T. Schoeman

Supervised by Professor Ryan M. Nefdt

A dissertation presented in partial fulfilment of the requirements for the degree of Master of Social Sciences in Philosophy

Faculty of Humanities

University of Cape Town

2024

Plagiarism Declaration

This research has not been previously submitted in whole, or in part, for the award of a degree. The research is my own and all significant contributions have been quoted, referenced, or attributed as necessary.

Signature:

Signed by candidate

Keldt Schoeman

Date: 24 July 2024

Acknowledgements

First, thank you to my supervisor, Ryan, who always knew when to reel me in from a freshly discovered tangent, but also gave me the space to pursue the project on my own terms. I have said as much to you, but I cannot think of a better person to have spent the last few years working with.

To my family, thank you for your patience, especially these last few months. I know it went on longer than expected, but finishing this project could not have happened without you. From being a sounding board, to talking me through the tough moments, to understanding the burning desire to do philosophy, I cannot adequately put my gratitude into words. Thank you.

I also owe my thanks to Damon, who never failed to inundate me with potential sources – some AI related, some corvid related, but none boring related. It is rare to find someone who can go from laughing about crows to earnestly talking about consciousness in just five minutes.

To Alex, who kept me surfing, and therefore sane in the darker moments. Your peer pressure seldom failed to get a laugh out of me, and it usually culminated with the both of us in the freezing Atlantic. Thank you for being a 'bad' influence.

Finally, Maike, thank you for always being willing to take time out of your busy schedule to read my work with me. Your attention to detail, generosity and willingness means a great deal to me generally, and to this project specifically.

Abstract

AI in the 21st century has come to be dominated by one school in particular, connectionism. And its successes are all around us – in the media we consume, in the music we listen to, in the cold calls we receive, etc. While this school was founded by psychologists, logicians, and philosophers with the goal of replicating human-level intelligence, the field has undergone a drastic transformation in recent years, entering a paradigm which is now dominated by engineering goals. Within this new paradigm, connectionism is no longer characterized as a field modelling the brain, but rather a mere engineering tool with incredible powers of pattern recognition. However, while the move to employ connectionist AI as a tool has led to remarkable successes in a variety of fields, it has also come with issues such as the black box problem, or epistemic opacity. Within a strictly engineering paradigm, attempts to explain the internal reasoning of these networks remain unsatisfying. Therefore, I propose recoupling connectionist networks with their roots in brain modelling, which would in turn open rich, new explanations for problems like epistemic opacity. Simply put, when we place the problem of opacity within the context of brain modelling, it appears that it may not be a problem at all, but an emergent feature of a complex system. In other words, we are beginning to have difficulty understanding modern connectionist networks in much the same manner we struggle to understand brains. Hence, it might well be feature, not a bug, that these systems should disappear into the mists of complexity.

Keywords: artificial intelligence, connectionism, classical computationalism, brains, epistemic opacity, emergence

Table of Contents

Table of Figures

Introduction

The overall focus of this research is on the problem of epistemic opacity. However, its aim is not to explain away opacity; various attempts to do this already exist within the literature. Instead, the aim of this project is to explain why epistemic opacity has come to exist in artificial neural networks (ANNs) at all.

The first chapter is concerned with the genealogy of modern ANNs and traces the history of these networks back to their connectionist roots. This history is divided into three epochs: brain modelling, revisionism and engineering. The first epoch, brain modelling, reviews the origins of both prepotent schools of AI, connectionism and classical computationalism. By tracing these schools, I argue two points: that connectionist networks, at least as Rosenblatt (1958) built them, began as models of brains; and that the criticisms which arose from the classical school, especially that of Minsky and Papert (1969), helped to decouple connectionist networks from these origins. The second epoch, revisionism, reviews the next major advance in the genealogy of ANNs, PDP networks, as well as the next wave of criticism related to them. Here I argue that the successful engineering of PDP networks (i.e., the use of new, highly effective mathematical techniques like backpropagation), as well as a fresh wave of criticisms, drove connectionist networks firmly into the engineering paradigm. Accordingly, the third epoch reviews recent ANN advances within this engineering paradigm such as deep learning, Transformers, etc. Here I highlight that the shift to the engineering paradigm has not only altered the personnel and methodology of the field, but also the kinds of conclusions which such research is supposed to be able to draw. My main argument is that while the success of modern ANNs did eventually blow past most criticisms from earlier epochs, these criticisms imparted an assumption - that the way ANNs have been engineered over recent decades has altered them so much that they are too different from brains to justify comparisons.

This assumption - that recent engineering advances have made ANNs so different from brains that comparisons cannot be justified - is investigated in the second chapter. I begin by reviewing some of the recent literature which, at least tentatively, draws comparisons between brains and ANNs. These comparisons are mostly in the domains of visual processing and language processing. Furthermore, they are not the basis for my argument, they merely indicate that comparisons are becoming more commonplace in the literature. To assess whether such comparisons are justified, I extend my discussion to Marr's three levels (1982) and use his

work to assess similarities. At the first level, that of hardware implementation, brains and ANNs have much in common. Both have local units, these are connected to one another, the connections are distributed across many local units, these local units therefore pass on information, connections between local units are weighted, these weightings can differ, and they can change. At Marr's second level, representation and algorithm, both brains and ANNs transform information from input to output in a remarkably similar way, using activation patterns. Moreover, the underlying principle these activation patterns take advantage of to process information is, in both cases, reliant on spatial relations to represent patterns of statistical dependency. At the third level, computational theory, we find yet more similarities. Not only are brains and ANNs similar insofar as they both have computational goals, but the kinds of computational goals which they excel at are also much the same. Most notably, these shared domains of functional competency are language processing and visual processing. Next, to explore what kind of comparison can be justified, I extend my discussion to the East Gate Centre, using it as evidence that ANNs can be compared to brains as a mechanistic abstraction of their spatial processing principles. Finally, I review the limitations of such a comparison.

Having established the historical link between brains and ANNs (the first chapter) and investigating how much similarity remains between them (the second chapter), the third chapter will argue that this connection with the brain gives us good reasons to view epistemic opacity in ANNs as an emergent feature, not an epistemological bug. To begin, I review definitions of emergence by Wimsatt (1997), Kim (1999) and Chalmers (2006) – using elements of each account to build a unified criteria. I argue that the criteria which best fit the referent for emergence involve both (1) difficulty providing satisfying scientific explanations, and (2) failures of aggregativity between constituents and high-level system properties. Next, I extend my discussion to epistemic opacity itself and argue that it fits both criteria. Following this, I review recent evidence from Hasson et al. (2020), which indicates, not only that brains and ANNs both leverage their processing with use of interpolation, but that ANNs have also converged on some of the same interpolative processing solutions which brains arrived at through evolution. Finally, I explore some potential objections to my argument, like inverted qualia, before attempting to reply to these objections.

1. On the Genealogy of Connectionism

In the history of AI, there are two dominant schools, one thought intelligence could be realised by a system of rules, the other thought it could be realised by learning. The rule-based school sought a system of symbols by which to represent human intelligence, while the learning-based school sought to literally replicate human intelligence by modelling the brain. In other words, one school viewed "computers as a system for manipulating mental symbols; the other, as a medium for modeling the brain. One sought to use computers to instantiate a formal representation of the world; the other, to simulate the interactions of neurons" (H. L. Dreyfus & Dreyfus, 1988, pp. 15–16). The rule-based, symbolist school has come to be known as classical computationalism. The learning-based, brain modelling school has come to be known as connectionism. The focus of this first chapter is the history of the brain modelling school, connectionism – a history which can be broken up into three epochs[1]. The principal aim of showing this history is to demonstrate how connectionist AI was decoupled from its roots in brain modelling and transformed into a mere engineering instrument.

1.1 The first epoch – brain modelling

1.1.1 Foundational breakthroughs

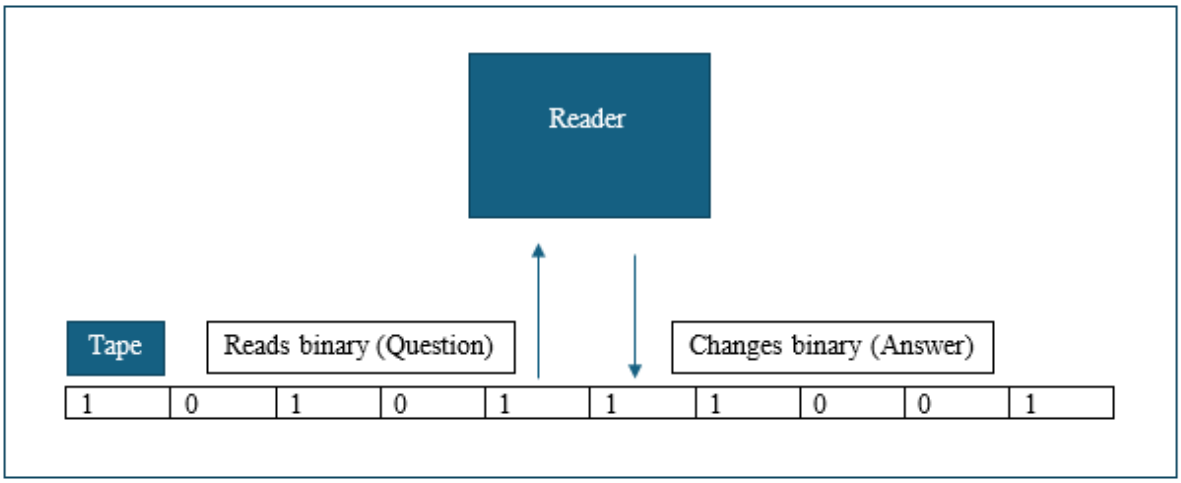
It may seem peculiar to begin a connectionist history with Alan Turing, the father of classical computationalism, however the antagonistic relationship between these two schools shaped them both. Therefore, no account of one is complete without some understanding of the other. In 1936, "Turing showed that every possible computation can in principle be performed by a mathematical system" (Boden, 2016, p. 8). Turing achieved this in his paper *On Computable Numbers, With An Application To The Entscheidungsproblem*, beginning with the claim that: "According to my definition, a number is computable if its decimal can be written down by a machine" (Turing, 1936, p. 1). Next, employing Leibniz's modern binary system[2] to represent all these potential computations, Turing is able to conclude "that all effectively calculable sequences are computable" (Turing, 1936, p. 34). This does exclude certain abstract areas of

[1] Although Pater does not distinguish the epochs as specifically as I do, his distinctions have played a role in my divisions here (Pater, 2018, p. 3).

[2] A system which was inspired by an ancient Chinese philosophical text, the Book of Changes. This book encoded Ying as a broken line and Yang as an unbroken line. When these lines were stacked in sixes on top of one another they would become one of sixty-four possible hexagrams. These hexagrams had the potential to represent any sequence of numbers or letters. Leibniz saw the potential in this system, and, substituting Ying and Yang for 1 and 0, he employed this system to create modern binary (Chalmers, 2022, p. 152).

mathematics, and, in 1936, the idea remained very much theoretical, but the hints of what such a machine could be capable of were evident even then. According to Boden, "[t]he core implication was clear … Turing computation could be applied to human and machine intelligence" (Boden, 2016, p. 10). Turing himself was very much aware of these implications. Part 6 of his paper is titled *The Universal Computing Machine* (Turing, 1936, p. 241) and he was of the position that its creation was a concrete possibility even then. Saying, "[i]t is possible to invent a single machine which can be used to compute any computable sequence" (Turing, 1936, p. 241). And the machine he envisioned all those years ago, came to look something like figure 1 below.

Figure 1: *An example of a Universal Computing Machine.*



The Universal Computing Machine would begin by reading the binary 1 or 0 on the tape. Then, based on its instructions, the reader would either change the binary (or not), and thereafter, the reader would either move left or right and repeat this same process on another piece of the tape. The original tape would be the 'question' and the transformed tape would be the 'answer'. In this way, Turing had created a machine which could theoretically compute any calculable sequence.

Before any practical advances on Turing's theory could come to fruition, another promising paper was published, and it theorised a different route to intelligence. Instead of creating a machine that could represent any computation, Warren McCulloch and Walter Pitts turned to something which was already doing such computations, the human brain. In their 1943 paper:

A Logical Calculus of The Ideas Immanent in Nervous Activity they laid the foundations for the theory of an ANN. Yet, the original purpose of their paper went further than just this claim, since, what they were really proposing was an explanation that neurons were some kind of proxy for a proposition, a sort of logical gate – clarifying that "[m]any years ago one of us, by considerations impertinent to this argument, was led to conceive of the response of any neuron as factually equivalent to a proposition" (McCulloch & Pitts, 1943, p. 117). Or as Boden puts it, "McCulloch and Pitts believed … that natural language boils down, in essence, to logic" (Boden, 2016, p. 10), and that the brain was the hardware which was responsible for instantiating it. Importantly, not only could this kind of claim explain the mechanisms behind human reasoning, but through this understanding, it also held the potential for the replication of such causal powers in a machine. The theory being, if we modelled the brain, we could replicate the properties of intelligence. Now, almost a century later, our best understanding of human brains and neural networks indicates that neurons do not necessarily represent a piece of information quite as large as a complete proposition. Rather, if there are representations of propositions occurring, current research indicates that it is better explained with reference to neural activation patterns. Yet, even if McCulloch and Pitts' claim may not have been borne out as true in the pure sense, much like Turing, their legacy has endured well into the 21st century. Specifically, their proposed notion that modelling the brain may be the most promising route to replicating human-level intelligence – since it gave rise to the brain modelling school of AI, connectionism.

As previously mentioned, the rationale for including the founding papers of both classical computationalism and connectionism is that the story of one cannot be told without the other. Together they represent the revitalization of an age old philosophical discord being taken up in the field of AI, that of the rationalist and the empiricist (Perconti & Plebe, 2020, p. 3). And this discord will come to direct connectionism in a crucial way, pushing this school away from its roots in brain modelling and towards the engineering paradigm; a paradigm which will come to view connectionist networks as no more than an efficacious pattern recognition instrument.

Therefore, in Turing's universal machine the rationalists found their rule-based answer for intelligence (classical computationalism), and in McCulloch and Pitt's theory of an artificial brain model empiricists found theirs (connectionism). The irony of this legacy is that McCulloch and Pitts' actual thesis proposed that neurons were a sort of logic gate which were supposed to represent propositions. Moreover, Pitts was a self-trained logician, and if he was to fall anywhere on this spectrum, he would likely be decidedly rationalist. Similarly, though,

while Turing is seen as the father of the rationalist school of AI, recent information has revealed that he was deeply sympathetic to the empiricist school and their ANNs. One of his early papers "even suggested computational approaches - such as neural networks and evolutionary computing - that became prominent only much later" (Boden, 2016, p. 9). This branch of Turing's work is most evident in his 1948 report for the National Physical Laboratory[3].

Simply put, Turing did not see the project of creating intelligence to be a purely rule based rationalist project and was more than just sympathetic towards the learning component already clear in human level intelligence, he was one of the pioneers of the field. Hence, he "accepted both goals of AI. [Turing] wanted the new machines to do useful things normally said to require intelligence … and also to model the processes occurring in biologically based minds" (Boden, 2016, p. 9). In this way, Turing was not a man driven by commitments to empiricism or rationalism, he was a man interested in unravelling the inner workings of intelligence.

My aim in drawing attention to this history is not to fan the flames of division further, particularly when it seems likely that both rationalist and empiricist elements are necessary for the creation of strong AI[4]. Instead, the aim is to point to the concrete effects this fracture has had on the development of Artificial Intelligence during the preceding century, particularly for connectionist AI. Before connectionism and classical computationalism became established schools, it was not uncommon for researchers from across different fields to employ both rationalist and empiricist elements in their attempts to understand intelligence. The examples we have already mentioned exemplify this. Pitts was a logician working on a theory of ANNs, while Turing was a mathematician and the father of classical computationalism, who was also working on a learning-based theory of intelligence. Which is not to say that the rationalist and

[3] In line with the stereotypical rationalist position one would usually expect him to hold, Turing does argue that "[m]any parts of a man's brain are definite nerve circuits required for quite definite purposes. Examples of these are centres which control respiration, sneezing, following moving objects with the eyes, etc." (Turing, 1948, p. 16). Which in turn implies that at least some elements of the brain are rule-based. However, simultaneously, he asserts that there are also "large parts of the brain, chiefly in the cortex, whose function is largely indeterminate" (Turing, 1948, p. 16) and this is where learning, a decidedly empiricist account of knowledge, enters his work. Making an analogy with infants and adults, Turing points out that these indeterminate parts do not yet have much of an impact on the infant but, later in "[t]he adult they have a great and purposive effect" (Turing, 1948, p. 16). The implication of which is rather simple: "the cortex of the infant is an unorganized machine which can be organized by suitable interfering training" (Turing, 1948, p. 16).

[4] Strong AI is the attempt to replicate the causal powers of the human mind in a machine. As Searle puts it, "according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states" (Searle, 1980, p. 417). Hence, strong AI is no less than the realisation of the ghost in the machine.

empiricist dogmas did not exist before the history of AI – they of course did; rather it is to say that once they began to dominate the AI landscape, attention which should have been broadly focused on the study of intelligence was focused on demonstrating that one or the other school was the genuine, scientific solution to the problem of intelligence. And, since this struggle for epistemic dominance has come to direct nearly a century of AI research, it is important that we are informed about the precise nature of its impact – that the acrimony between the two schools decoupled connectionist AI from its roots in modelling the brain.

Turing would live long enough to help design the first computer in 1948, but not long enough to work with its more powerful descendants, or its connectionist cousins (Boden, 2016, p. 8). This meant that, since his 1948 report for the National Physical Laboratory did not see the light of day until 1994 (Perconti & Plebe, 2020, p. 3), the AI community remained unaware that Turing "was the first to advance the idea that computers can be designed simply by letting them learn by themselves. [And that] [h]e [even] envisioned a machine based on distributed interconnected elements, called *B-type unorganized machine*" (Perconti & Plebe, 2020, p. 3). In the absence of such critical information, it would be the rationalist school of artificial intelligence, classical computationalism, which would largely dominate the following decades of AI research.

1.1.2 Notable early developments

A curious exception to this trend was Marvin Minsky. The project he chose to work on for his PhD was the creation of a connectionist computer called SNARC (Stochastic Neural Analog Reinforcement Computer). With its completion in 1954, it was to be "the first neural computer, assembling 40 'neurons', each made with six vacuum tubes and a motor to adjust its connections mechanically" (Perconti & Plebe, 2020, p. 3). The project was not the sensational media success which came to be associated with later connectionist networks, but the attempt to create the first neural computer by modelling the brain on a small scale, was certainly ambitious. If one is aware of the significant role Minsky played in the history of AI, the fact that his PhD was connectionist in nature may appear odd, however, in the following pages I will argue that the opposite is actually the case; since, being an expert on connectionist networks meant Minsky was perfectly placed to criticize them.

One of the next major developments in 1950s AI research was by Allan Newell and Herbert Simon. The two men theorised about the creation of a Logic Theory Machine. At the time of their first paper it did not even exist as a computer, but had to be simulated by hand (Newell &

Simon, 1956, p. 79). Within just a few months they recruited Cliff Shaw to turn their idea into reality. The result being that "The Logic Theorist (LT) of Newell, Simon, and Shaw, [was] presented at the inaugural 1956 AI conference at Dartmouth, [and] managed to prove thirty-eight out of the fifty-two propositional-logic theorems of Principia Mathematica" (Arkoudas & Bringsjord, 2014, p. 37). It even "found a more elegant proof of one of them" (Boden, 2016, p. 11). Russell, one of the authors of *Principia Mathematica*, "was delighted by this achievement, but the Journal of Symbolic Logic refused to publish a paper with a computer program named as an author, especially as it hadn't proved a new theorem" (Boden, 2016, p. 11). This success bolstered rationalist convictions about AI. Not only had the Logic Theory Machine proved the existing theorems of some of our foremost logicians, but it had also been able to improve upon one of them by employing rule-based reasoning. And here was perhaps the first real suggestion that one day a computer may well become more intelligent than ourselves.

Newell, Simon, and Shaw would continue to work on their idea, later turning it into a program with broader applications than the Logic Theory Machine (Boden, 2016, pp. 11 & 12). Appropriately this new program would be called the General Problem Solver (GPS), and they would go on refining it well into the 1960s[5]. With this loose structure the GPS experienced quite a few high-profile successes[6]. More importantly however, what it represented was a new way a machine could be considered intelligent. The GPS was not necessarily better than a human at any single task, but it performed well across a wide range of problems. This idea of general intelligence was not as common in the 1960s as it is today. Hence, the GPS was important, not only for inspiring work on the distinction between general intelligence and narrow intelligence, but also for being one of the early intimations that Artificial General Intelligence (AGI) was possible in practice. Considering this, and that much like the Logic

[5] The General Problem Solver, or the GPS for short, had far broader applications than the Logic Theory Machine. As Boden explains, the "GPS could be applied to any problem that could be represented … in terms of goals, sub-goals, actions, and operators. It was up to the programmers to identify the goals, actions, and operators relevant for any specific field. But once that had been done, the reasoning could be left to the program" (Boden, 2016, p. 12).

[6] Perhaps the most famous success was when the GPS managed to solve the problem of the missionaries and cannibals (Boden, 2016, p. 12), a problem humans struggle with. The problem is as follows: there are 3 missionaries and 3 cannibals who need to cross a river. If the cannibals ever outnumber missionaries, the missionaries will be eaten. There is a single boat to cross the river, but it can only carry two at a time. How do you get all 6 individuals across the river safely?

Theory Machine, the GPS was firmly rationalist in design, its successes further bolstered convictions that rationalism was the route to the realization of artificial intelligence.

The idea that computers could one day become more intelligent than ourselves was to be reiterated just a few years later by the work of Arthur Samuel. Samuel created a checkers playing computer which "made newspaper headlines because it learned to beat Samuel himself" (Boden, 2016, p. 11). In his own paper on the matter Samuel concludes that "one can say with some certainty that it is now possible to devise learning schemes which will greatly outperform an average person" (Samuel, 1959, p. 223). Given the vernacular about learning, one may be forgiven for thinking this computer fits with the connectionist school of AI. However, that is not the case and Samuel's computer was of the classical kind[7], serving as more evidence that the route to intelligence was the one envisioned by the rationalist.

The next example is a particularly unusual one in the history of AI. It is best known as Pandemonium[8]. Originally presented at a symposium in late 1958, it was a computer with a focus on learning. Boden considers it to be the instantiation of a parallel processing system using classical computationalism. Yet, the presence of classical and connectionist elements makes Pandemonium incredibly difficult to place on the side of rationalism or empiricism. There are rationalist elements that employ formal logic, rules and binary code, while there is also a learning architecture which bears stark resemblance to the brain models used to instantiate empiricist connectionist networks (Selfridge, 1959, pp. 15–22). However, regardless of which side this project should be assigned to, one thing is clear, the boundaries between the two schools of AI were not yet concrete in 1958. Considering this - another example difficult

[7] What Samuel's created was a precursor to more advanced types of classical computationalism - like that of the chess computer Deep Blue II. Deep Blue II was a rule-based system of the 1990s which was able to defeat the chess world champion Gary Kasparov. However, the technology at the foundation of its success was of the classical kind. And Samuel's computer was just a rudimentary form of Deep Blue II. Where Deep Blue II had a "massively parallel system designed for carrying out chess game tree searches" (Campbell et al., 2002, p. 60), Samuel's checkers computer carried out a much smaller scale of game tree searches, and carried them out for a different game, checkers. Hence, the major difference between the two was that of scale and game type. Samuel's computer could only explore a very limited sequence of moves in its checkers search tree (Samuel, 1959, pp. 213–217), whereas Deep Blue II could search up to 330 million chess positions in a second. Yet, what is common to both computers is that they still relied on the rationalist if-then rules of classical computationalism.

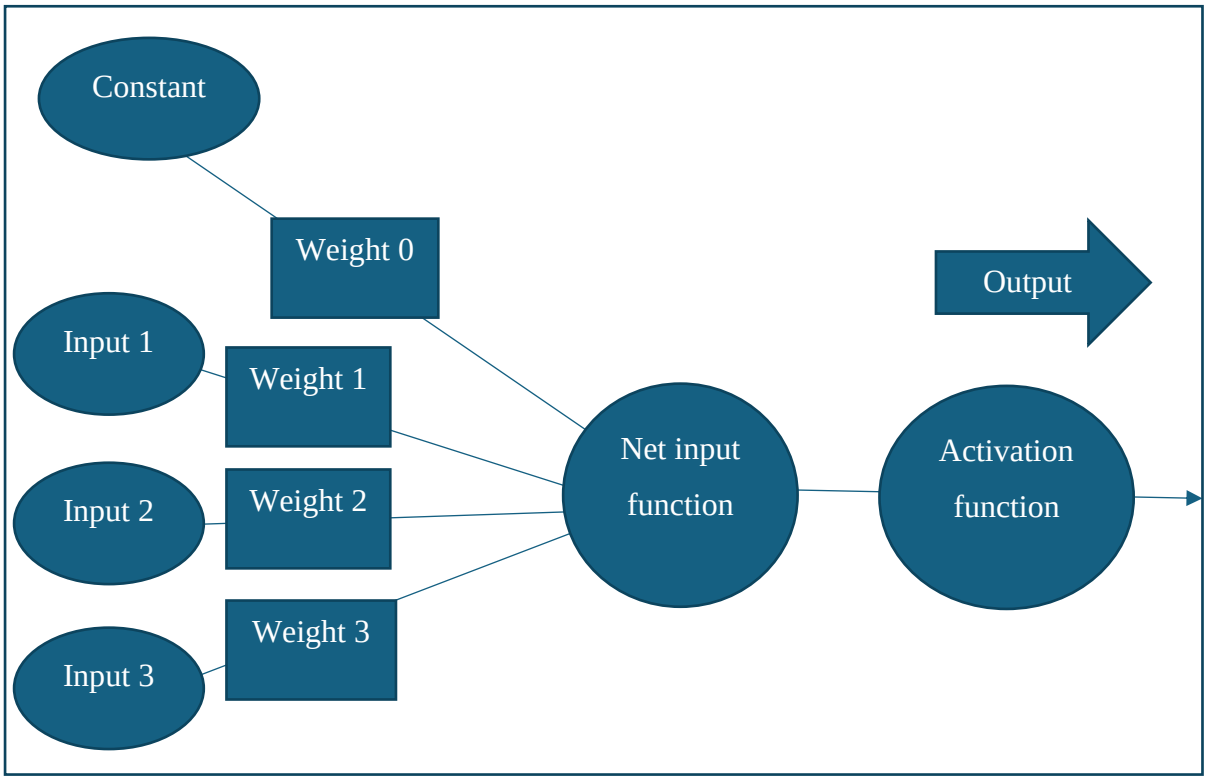
[8] Pandemonium "learned to recognize patterns by having many bottom-level 'demons', each always looking out for one simple perceptual input, which passed their results on to higher level demons. These weighed the features … downplaying any features that didn't fit. Confidence levels could vary, and they mattered: the demons that shouted loudest had the greatest effect. Finally, a master-demon chose the most plausible pattern, given the (often conflicting) evidence available. This research soon influenced both connectionism and symbolic AI" (Boden, 2016, pp. 15 & 16).

to call either empiricist or rationalist - I will not be employing the terms rationalist and empiricist from this point onwards. Although they are useful in capturing the context of the underlying discord, having done so, they have served their purpose and further use would only confuse matters.

1.1.3 The Perceptron

While classical computationalism had taken centre-stage during the 1950s, the work of Frank Rosenblatt was about to challenge previous approaches and place the connectionist school of AI in the spotlight. For, while the Perceptron project was simply "a photoelectric machine with eight 'neurons' and connections that can be adjusted according to a learning rule" (Perconti & Plebe, 2020, p. 3), the pattern recognition abilities it demonstrated were remarkable. Below is figure 2, a basic diagram of the original Perceptron.

Figure 2: *The structure of the Perceptron.*



At its inception, the Perceptron was entirely mechanical, and the inputs were hooked up to photocells. Once the input from the photocells entered the network, the weighted input data would give a net function of the entire system as an output. In other words, Rosenblatt's Perceptron was receiving information from the external world via the input photocells and producing a function as an output. In the case of the Perceptron, it was taking visual information

from the photocells, processing it, and outputting it as a function which could be used to recognize shapes. The media coverage of the project was vigorous and a newsletter from *Science* went as far as to go with the title *Perceptron Thinks*, saying that it "literally teaches itself to recognize objects the first time it encounters them" (The Science News, 1958, p. 39). While this may have been a sensationalist narrative, Rosenblatt himself was quietly optimistic, asserting that "[b]y the study of systems such as the perceptron, it is hoped that those fundamental laws of organization which are common to all information handling systems, machines and men included, may eventually be understood" (Rosenblatt, 1958, pp. 407 & 408). In other words, in our attempts to create successful imitations of the properties possessed by the human brain, we may one day unravel the mystery of intelligence too. However, as promising as its successes were, the Perceptron itself was still very limited in its applications. After all, it was just a rudimentary model of parallel processing in the brain which had been applied to the problem of vision. Yet, despite these limitations, Rosenblatt was adamant that since the technology was based on the human brain, the applications would come to be similarly broad. A few years later in 1962 he went as far as to say that:

> [a] perceptron is first and foremost a brain model, not an invention for pattern recognition. As a brain model, its utility is in enabling us to determine the physical conditions for the emergence of various psychological properties. It is by no means a "complete" model, and we are fully aware of the simplifications that have been made from biological systems; but it is, at least, an analyzable model (Rosenblatt, 1962, p. vi).

This makes clear that Rosenblatt identified connectionist technology like the Perceptron not just as some instrument for pattern recognition, but as a model of the brain - a view which is radically different from the one which has become dominant today. In subsequent years, connectionist technology has shifted towards engineering goals (Perconti & Plebe, 2020, p. 1) and taken on the view which Rosenblatt rejected, that such models are just highly effective tools for pattern recognition. This shift did not take place overnight, it took decades of innovation and multiple waves of criticism. The rest of the chapter will give an account of just how connectionism transformed from the brain modelling school of AI we find in the 1960s to the engineering paradigm we find today.

1.1.4 Initial criticism

The origins of this decoupling begin, in part, as the result of the success of Rosenblatt's work. The promise connectionist networks had shown by modelling the brain posed a threat to those committed to the idea that classical computationalism was the answer to intelligence. In this way, Rosenblatt's Perceptron would prove to be a key moment in the fractious relationship between connectionism and classical computationalism. Boden summarizes this escalation:

> The bad feeling on the … connectionist side began as a mixture of professional jealousy and righteous indignation … Members of the symbolist camp were initially less hostile, because they saw themselves as winning the AI competition. Indeed, they largely ignored the early network research … In 1958, however, an ambitious theory of neurodynamics … was presented by Frank Rosenblatt and partially implemented in his photoelectric Perceptron machine … This novel form of connectionism couldn't be ignored by the symbolists (Boden, 2016, pp. 18 & 19).

The details really matter here, particularly regarding who was in attendance when Rosenblatt showed his preliminary findings. For, "Rosenblatt presented an early version of his perceptron research at MIT to an AI group in the fall of 1958 … Amongst the members of that audience was Minsky, Rosenblatt's high school colleague, and the future first author of the critical appraisal of perceptrons" (Pater, 2018, p. 3). However, this was no longer the same Minsky who had done a connectionist PhD. In the years following his doctoral thesis, Minsky had a change of heart and was now firmly rooted on the side of classical computationalism. Accordingly, having been trained in connectionist architecture, having switched allegiances, having early access to Rosenblatt's research, and then seeing how enthusiastically the press responded to the Perceptron, Minsky was perfectly placed, and sufficiently motivated to launch a devastating critique. In this way it would be the success of Rosenblatt's work, along with the powerful claims that came with it, which would endanger the future of connectionist research. And, although it would be years before the criticism began to take any formal shape, the delay would not matter.

> [From] [a]bout 1965, Minsky and Papert, who were running a laboratory at MIT dedicated to the symbol-manipulation approach and therefore competing for support with the perceptron projects, began circulating drafts of a book attacking the idea of the perceptron (Dreyfus & Dreyfus, 1988, p. 21).

While the completion of the book would take a few more years, by 1969, Minsky and Papert, were ready to show the world exactly why the applications of Rosenblatt's work were limited enough to be considered a dead end. In their own words, they explain that the project began because they were "[a]ppalled at the persistent influence of perceptrons (and similar ways of thinking) on practical pattern recognition, [and thus] we were determined to set out our work as a book" (Minsky & Papert, 1969, p. 242). As mentioned earlier in the chapter, one of these men, Minsky, had begun his career in the connectionist school of AI, building the first ever neural computer, SNARC. It was not quite the success that Rosenblatt's Perceptron was, but this piece of information is crucial because Minsky's conversion to the side of classical computationalism would now lend weight to any criticisms he made. With his name on the book, *Perceptrons* became, not simply an entire book dedicated to outlining the limits of connectionism, but one written by a man with significant expertise in connectionist research. Consequently, it was even more damning for connectionism, and "achieved the desired effect, marginalizing artificial neural network research for decades" (Perconti & Plebe, 2020, p. 3). As Dreyfus and Dreyfus put it,

> Rosenblatt was discredited along with the hundreds of less responsible network research groups that his work had encouraged. His research money dried up, and he had trouble getting his work published. By 1970, as far as AI was concerned, neural nets were dead (Dreyfus & Dreyfus, 1988, pp. 23 & 24).

And it is for this reason that we have needed to follow both the history of connectionism and classical computationalism. Without both sides of this history, it is impossible to understand why connectionist research disappeared for two decades, and why, later, it would be transformed into an engineering paradigm.

1.2 The second epoch - revisionism[9]

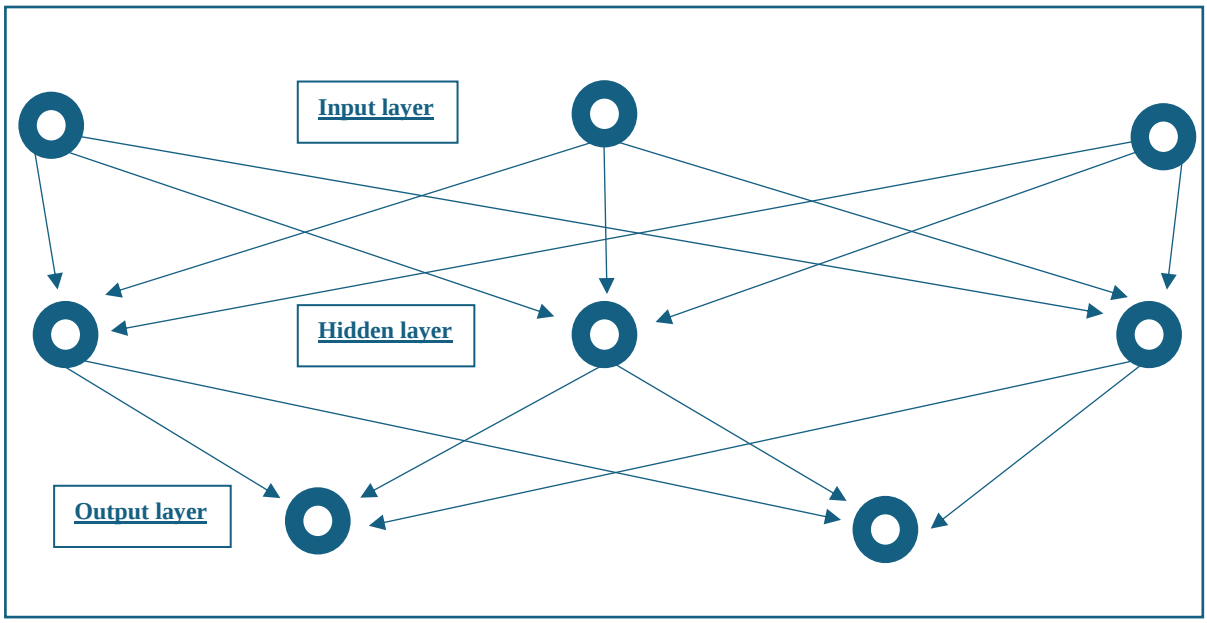
1.2.1 PDP networks

In hindsight, the enthusiasm with which the AI community relegated connectionist research was overzealous. Minsky and Papert had compared the greatest successes of classical computationalism to the weakest possible projections of what connectionism maybe capable

[9] I found the revisionist label for this epoch in Pinker and Prince's (1988) criticism. They mention it in the context of the past tense debate but employ it to characterize the general transformation of connectionist networks brought about by the work of Rumelhart and McClelland.

of. While Rosenblatt suspected that as network complexity grew, so would network capabilities, Minsky and Papert disagreed. Their intuition was that the properties of single layer connectionist networks would not scale up as they grew into multi-layered networks (Boden, 2016, pp. 92 & 93). It would take 17 years, but Rosenblatt would be proven right[10]. In 1986, Rumelhart and McClelland, along with a young Geoffrey Hinton, published *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* - proving that as network complexity grew, so could network capabilities. This generated renewed interest in the field and effectively ended the AI winter for connectionism. Below, figure 3, is a basic diagram of a PDP network.

***Figure 3**: The basic structure of a PDP network.*



In reality, these networks are more complex and much larger in terms of parameters than the above figure, but the simplification allows us to highlight how PDP networks differed from the Perceptron. The basic principle remains the same – data is fed into the network as an input and the result is the generation of a function as an output. However, in the case of PDP networks, each artificial neuron is connected to every other artificial neuron in the subsequent layer – which means that what went into the network as a whole piece of information is broken up and processed in parallel to generate the activation function. Crucial to this kind of parallel

[10] Sadly, Rosenblatt would not live long enough to see his work vindicated. Instead, he would die in a tragic accident in 1971, just two years after the publication of Perceptrons (Pater, 2018, p. 1).

processing was the inclusion of a hidden layer and a far greater abundance of connections. Yet, these structural changes would have been far less effective without the final difference - PDP networks employed a training technique called backpropagation[11]. This technique is best explained by comparing it to Rosenblatt's Perceptron. Both networks would compare the actual results with the desired results and adjust the weights to optimize the function. With the Perceptron, researchers would manually play around with the weights, slowly improving the results, while with PDP networks backpropagation meant that they could automatically compare the actual output and desired output with a mathematical rule determining how to alter the weights to reduce error.

With the success of PDP networks, stories about the early connectionist critics would also begin to surface, and it would come to light that Minsky and Papert's "initial critique had dripped with vitriol…[and that] [t]he draft was even more venomous: friendly colleagues persuaded them to tone it down, to give the scientific points more prominence." (Boden, 2016, p. 94). Considering this, and that funding for connectionist research all but dried up after their criticism, a common insinuation is that Minsky and Papert's motivation for publishing their book was to secure control of AI funding (Boden, 2016, p. 94). Especially since the criticism had conveniently funnelled most of the AI funding into their own field, classical computationalism. Consequently, once PDP networks contradicted their predictions for connectionism, one may expect that Minsky and Papert would revisit their position, but "they were unrepentant…they insisted that high-level intelligence cannot arise from pure randomness… [and] [t]hey protested that their critique wasn't the only factor that had led ANNs into their wilderness years" (Boden, 2016, p. 94). Elements of this response hold some truth but for the connectionist researchers who had been stuck on the wrong side of funding for all those years, such guarded platitudes were little consolation.

Perhaps the most ironic part of this history is that Minsky and Papert succeeded in accomplishing the opposite of what they had intended. Although they delayed connectionist advances, once Rumelhart and McClelland published their book, public opinion swung drastically against the classical school of AI. This was by no means what Rumelhart and McClelland had intended either. Sensitive to the conflict which had relegated their own research to the periphery, they argued that any successful attempt to create artificial intelligence

[11] Back propagation refers to a mathematical technique to improve network function. The network runs forward to find the activation function (as it normally would), except now it also runs in reverse, adjusting the weights by using a technique first described by Paul Werbos in his PhD thesis. However, in 1974 it was not called backpropagation, Werbos described it as an "algorithm of dynamic feedback" (Werbos, 1974, p. 21).

would require both schools (Boden, 2016, p. 93). However, problematically, classical computationalism had been sold to the public as a silver bullet, and it had been given almost two decades of exclusive funding to produce results. In the absence of those results, it began to look like the dead end which it had promised connectionism to be. Moreover, technical and philosophical problems had begun to emerge around this time (Arkoudas & Bringsjord, 2014, p. 50). The most famous of which were Searle's Chinese Room Argument[12], Ned Block's Chinese Nation Thought Experiment[13], and Hubert Dreyfus' criticisms in *What Computers Can't Do*[14]. Besides these problems, the PDP networks of Rumelhart and McClelland were yielding increasingly significant results. Not only had technology improved to meet the demands of these networks, but the work of Rumelhart and McClelland had begun to employ the kind of mathematics, which was generally used to solve engineering problems. For, as mentioned earlier, in the absence of proper funding, "[t]he success of PDP was largely due to an efficient mathematical rule, known as backpropagation, for adapting the connections between units, from examples of the desired function between known input and output" (Perconti & Plebe, 2020, p. 3).

This is an important turning point for connectionist AI. Until this moment, ANNs had largely been a study of cognition which had its roots in brain modelling. Consequently, most of the researchers working in connectionism were psychologists. For example, McCulloch was one, as was Rosenblatt and most of the main members from the PDP group - McClelland,

[12] The Chinese Room Argument places a non-Chinese speaker in a room alone with a book containing answers for every possible Chinese question. Chinese questions are passed under the door and the non-Chinese speaker must match the question symbols with the corresponding answer symbols, sliding the answer back under the door. This undermines classical computationalism on the basis that such pure symbol manipulation will never be enough to constitute understanding, since the non-Chinese speaker may appear to understand the questions without even understanding the Chinese language (Searle, 1980).

[13] The Chinese Nation thought experiment envisions a hypothetical scenario in which consciousness is instantiated by the nation of China. Each citizen receives instruction via a two-way-radio and the combination of these connections (and the symbol manipulation that their interactions represent) is supposed to result in the instantiation of consciousness for the classical computationalists. For example, if citizen b receives instruction 1 in state x, they are to perform action y. The problem this raises is that creating a system of classical computationalism which is only functionally equivalent to a brain through symbol manipulation raises serious doubts about the presence of qualia in this system. In other words, is there something that it is like to be the Chinese nation? (Block, 1978).

[14] Dreyfus' criticisms are perhaps the most powerful of the three. He argues that classical computationalism is altogether the wrong kind of computer to realize the goal of strong AI. For, considering the vagueness and complexity of the world which is supposed to have brought about human level intelligence, the strict rule-based system of symbols will always be inadequate. Instead, the kind of machine that could do natural language, visual processing et cetera. would need to be embodied and "exhibit the forms of 'information processing' essential in dealing with our nonformal world" (H. Dreyfus, 1972, p. 216).

Rumelhart, Hinton, etc. (Perconti & Plebe, 2020, p. 1). Pitts was an exception in that he was a self-trained logician. Yet, he would not remain an exception for much longer. With the success of backpropagation, "AI largely changed direction in the 1980s and 1990s, concentrating on building domain specific systems and on sub-goals such as self-organization, self-repair, and reliability. Computer scientists aimed to construct 'intelligence amplifiers' for human beings, rather than imitation humans" (Proudfoot & Copeland, 2012, p. 2). In other words, with the publication of Rumelhart and McClelland's book, the focus of connectionism had begun to shift. Where it had once been a domain driven by psychologists attempting to understand the inner workings of intelligence by modelling the brain, it would grow to become a project of engineering and optimization. Or as Perconti and Plebe put it:

> one of the most distinctive differences between the first generation of artificial neural networks and the current deep learning enterprise … [is] its focus. The primary motivation for the development of the early neural networks was the study of cognition… [whereas now] the scope has drastically shifted towards engineering goals (Perconti & Plebe, 2020, p. 1).

And there appear to be two clear reasons for this shift, both of which stem from conflict with classical computationalism. Firstly, the absence of proper funding had led Rumelhart, McClelland and their PDP research group to find other ways around the need for raw computing power. This in turn led them to find engineering solutions to the problems their networks encountered, an approach those that followed them would be inspired by. Secondly, connectionisms close relation with strong AI had nearly destroyed the entire field. It was only because of a few dedicated researchers working contrary to funding that the field survived. Hence, while unravelling components of intelligence (such as visual processing) seemed like a realizable goal, claims about the creation of strong AI would likely lead to criticisms which would only damage the revival of the field. Therefore, moving towards engineering goals also served as a pragmatic response to the problem posed by the potential for another AI winter. It is for this reason that I characterize the second epoch of connectionist AI's history as a revisionist one. The challenges posed by the first AI winter inspired a revision in methodology towards the novel engineering solutions far more commonplace today.

Yet, this revision in methodology was not a magical shift motivated by some simultaneous, realization about the power of engineering. Instead, there were serious sociological factors which influenced the move. Since, every major advance in connectionism came along with

strong criticisms about the limitations of connectionist networks, and this epoch was no different. Perhaps the two most famous of these criticisms come from Pinker and Prince as well as Fodor and Pylyshyn. Unlike with Rosenblatt, the criticisms to Rumelhart and McClelland's work did not take 11 years to formulate. Pinker and Prince had a criticism ready within less than two.

1.2.2 Pinker and Prince criticism

Influenced by the idea of a Universal Grammar[15] (UG), Pinker and Prince's broader criticism was that:

> Rumelhart and McClelland have described a connectionist … model of acquisition … yet the model contains no explicit rules, only a set of neuron-style units … [and they] conclude that linguistic rules may be merely approximate fictions and that the real causal processes in language use and acquisition must be characterized as the transfer of activation levels among units and modification of the weights and their connections … We conclude that connectionists' claims about the dispensability of rules in explanations in the psychology of language must be rejected, and that, on the contrary, the linguistic and developmental facts provide good evidence for such rules (Pinker & Prince, 1988, pp. 73 & 74).

The best evidence they provide for the importance of these rules in language is to be found in their criticism of the way connectionist networks handle the transformation of a word to the past tense. They argue that not only do these networks get a substantial number of these transformations wrong, but in getting them wrong they can "acquire rules that are not found in any language" (Pinker & Prince, 1988, p. 180). It is these mistakes which indicate to Pinker and Prince that the rule-based account of language is the proper one. While these mistakes may give some evidence to side with the rule-based account of language, Pinker and Prince's argument goes further, claiming that even if there were a connectionist network in the future without such mistakes, the rule-based account would still be the correct one. For, if the connectionist is

> to retreat from the claim that the RM model in its current form is to be taken as a literal model of inflection acquisition … [rather arguing that] these problems would all diminish if more sophisticated kinds of PDP networks were used. … a successful PDP

[15] Universal Grammer, refers to the idea that below the messy surface of every language is the same essential structure regardless of whether the language is English, Chinese or Zulu (Chomsky, 1965, p. 6).

> model of more complex design may be nothing more than an implementation of a symbolic rule-based account … [And] there is no basis for the belief that connectionism will dissolve the difficult puzzles of language, or even provide radically new solutions to them … [Thus,] when the deeper and more diagnostic patterns are examined with care, one sees not only that the PDP model is not a viable alternative to symbolic theories, but that the symbolic account is supported in virtually every aspect (Pinker & Prince, 1988, pp. 181–184).

Simply put, Pinker and Prince's problem with connectionist networks is not just that they make mistakes, but that even if they did not, they would simply be modelling the rule-based account of language[16]. The context here is that Pinker and Prince, both of whom are heavily influenced by the work of Noam Chomsky on generative linguistics, are already convinced that there is such a thing as a UG. This means that any theory of language which can account for acquisition without an explicit rule-based framework, such as connectionism supposedly can, is simply mistaken about what is really going on. For, in that case, the real causal process is not the activation patterns, but the rules which the activation pattern are able to map. Therefore, even if connectionist networks get the technical elements right, the central thesis that language is the pure result of learning will remain incorrect because what the connectionist networks are successfully modelling is, in fact, a rule-based system of language. With all that being said, contemporary literature as well as earlier responses like that of Elman (1990), suggest that the most accurate account of language may well be a balance between both positions. Especially since the poverty of stimulus argument advanced by the UG framework raises serious questions about innateness - questions which remain largely unanswered.

1.2.3 Fodor and Pylyshyn criticism

The next criticism comes from Fodor and Pylyshyn. Their issue with connectionist networks is, in some ways, analogous to the previous criticism insofar as there is an assumption of a rule-based system present in both. It is in these similarities that we find evidence for the battle between connectionism and classical computationalism once more. However, the point Fodor and Pylyshyn are making is the following:

> discussions of the relative merits of the two architectures have thus far been marked by a variety of confusions and irrelevances. It's our view that when you clear away these

[16] Which is especially interesting given that recent literature argues the opposite, making the claim that the symbolic rule-based accounts are a subset of the statistical deep learning ones (Piantadosi, 2023).

> misconceptions what's left is a real disagreement about the nature of mental processes and mental representations. But it seems to us that it is a matter that was substantially put to rest about thirty years ago; and the arguments that then appeared to militate decisively in favor of the Classical view appear to us to do so still … We claim that the major distinction is that, while both Connectionist and Classical architectures postulate representational mental states, the latter but not the former are committed to a symbol-level of representation, or to a 'language of thought': i.e., to representational states that have combinatorial syntactic and semantic structure (J. Fodor & Pylyshyn, 1988, pp. 1–4).

In other words, the problem with connectionism is that the best we can hope for from a connectionist network, at least in terms of representation, is that it will be able to accomplish what may be called sub-symbolic representations (J. Fodor & Pylyshyn, 1988, pp. 5 & 6). While the architecture of classical computationalism enables representations which go beyond mere sub-symbolic representations and is capable of what Fodor and Pylyshyn call a combinatorial syntactic and semantic structure in their representations.

Put another way, because connectionist networks handle information in a distributed fashion, no single artificial neuron represents a single piece of information. Rather, the information is spread out across the network. For example, if we are using a network to process the common features of dogs, there will be no single location we can point to and say, 'that represents the tail of a dog'. The tail of a dog is represented in a distributed way which is analogous to a neural activation pattern. Therefore, the features represented by any single artificial neuron can, in a somewhat rugged analogy, be understood as sub-symbolic. It is only when we step back and see the forest for the trees, when we look at the larger activation pattern, that we can begin to point to some kind of symbol representation which is present in the network. Hence, representations in these networks are somewhat controversially accepted as sub-symbolic[17]. Fodor and Pylyshyn are willing to accept this sub-symbolic explanation of representations in connectionist networks, but their criticism is that the networks can go no further. They argue that the very architecture of the connectionist network is not committed to a symbol-level of representation because the architecture itself lacks the necessary combinatorial syntactic and semantic structure. They also argue that the relationships between the symbols, the combinatorial syntactic and semantic structures which link the dog to the tail, cannot, by the

[17] See the Harmonic Mind (Smolensky & Legendre, 2006), where Smolensky and Legendre challenge this notion by merging the two into a neuro-symbolic model.

very nature of the connectionist architecture, be said to be present. The examples which they often use to demonstrate these points are ‘John loves Mary’ and ‘a cup has a handle’. Arguing that:

> [r]eal constituency does have to do with parts and wholes; the symbol ‘Mary’ is literally a part of the symbol ‘John loves Mary’. It is because their symbols enter into real-constituency relations that natural languages have both atomic symbols and complex ones. By contrast, the definition relation can hold in a language where all the symbols are syntactically atomic; e.g. a language which contains both ‘cup’ and ‘has-a-handle’ as primitive predicates. This point is worth stressing. The question whether a representational system has real constituency is independent of the question of microfeature analysis; it arises both for systems in which you have CUP as semantically primitive, and for systems in which the semantic primitives are things like ‘+ has-a-handle’ and CUP and the like are defined in terms of these primitives. It really is very important not to confuse the semantic distinction between primitive expressions and defined expressions with the syntactic distinction between atomic symbols and complex symbols (J. Fodor & Pylyshyn, 1988, p. 13).

Simply put, in their view, connectionist networks are only doing the one kind of representation – the atomic kind which is a primitive sort of microfeature analysis. The larger structure of real constituency (complex rule-based relations between symbols) is distinct from the sub-symbolic kind and is not found in connectionist networks – which leads to the conclusion that these networks are not committed to a ‘language of thought’. It is at this point that we can see Fodor and Pylyshyn’s criticism beginning to resemble that of Pinker and Prince’s. However, the claim they are making is even stronger. Pinker and Prince were willing to concede that, in the event of a more sophisticated connectionist network succeeding in modelling language, it would simply be the instantiation of a symbolic rule-based account. Fodor and Pylyshyn disagree; arguing that the architecture of connectionist networks is, by its very nature, incapable of modelling language at the level of combinatorial syntactic and semantic structure - that it is not capable of a language of thought. Whether this claim has stood the test of time, especially considering the successes of contemporary generative language models, remains a hot button issue.

1.2.4 Responses to criticisms

Jeffrey Elman was one of the first to identify some of the problems with these criticisms. In his own words, he explains that:

> the most fundamental concepts of linguistic analysis have a fluidity, which at the very least, suggests an important role for learning; and the exact form of the those concepts remains an open and important question. In PDP networks, representational form and representational content often can be learned simultaneously. Moreover, the representations which result have many of the flexible and graded characteristics noted above (Elman, 1990, pp. 191 & 192).

This position openly contradicts the main criticism of Fodor and Pylyshyn. If representational form and content can be learned simultaneously, then the idea that connectionist networks can only ever succeed at representation of the sub-symbolic demonstrates a misunderstanding of what these networks are doing. Elman even goes so far as to explicitly mention Fodor and Pylyshyn, as well as their claim (Elman, 1990, p. 203). He then goes on to substantiate his position with reference to the success of connectionist models at predicting the next class of word, explaining that

> individual items are typically not very predictable but classes of words are. This is precisely the pattern found here, in which the error in predicting the actual next word in a given context remains high, but the network is able to predict the approximate likelihood of occurrence of classes of words (Elman, 1990, p. 202).

At the very least, this appears to be some form of combinatorial syntactic structure present in connectionist networks. The ability to predict where verbs, nouns, adjectives, etc. go is an example of basic syntactic proficiency. Though, Elman is not satisfied with demonstrating just this, he goes on to explain that:

> [a] finer grained analysis reveals that the network also distinguishes between the specific occurrences of each lexical item, that is, the tokens. The internal representations of the various tokens of a lexical type are very similar. Hence, they are all gathered under a single branch in the tree. However, the internal representations also make subtle distinctions between (for example), boy in one context and boy in another … [and] it is useful to try to understand these results in geometric terms. The hidden unit activation patterns pick out points in a high (but fixed) dimensional space. This is

> the space available to the network for its internal representations. The network structures that space in such a way that important relations between entities is translated into spatial relationships. Entities which are nouns are located in one region of space and verbs in another. In a similar manner, different types (here, lexical items) are distinguished from one another by occupying different regions of space; but also, tokens of a same type are differentiated. The differentiation is non-random, and the way in which tokens of one type are elaborated is similar to elaboration of another type. That is, John1 bears the same spatial relationship to John 2 as Mary 1 bears to Mary 2 (Elman, 1990, pp. 205–207).

The reference to John and Mary appears intentional, because, although these are common example names, they are also the precise names chosen by Fodor and Pylyshyn in their criticism. The point Elman is making is that if these networks can distinguish different contextual usages which involve 'John' and 'Mary', such as 'John loves Mary' or 'Mary loves John', they are representing the combinatorial semantic structure spatially, not just combinatorial syntactic structure. This undermines Fodor and Pylyshyn's core criticism. As for Pinker and Prince, their criticism is not as strong as Fodor and Pylyshyn's, and as a result, it is not as easily dismissed. Yet, as Elman points out, whether representation is learning based, rule based, or both, remains an open question. Since, at the time of writing the paper, there was no clear evidence to exclude any of these options. Consequently, while we cannot preclude the possibility that any connectionist network which successfully models language is merely modelling the rule-based system, the claim that this is definitively what is happening in a successful network is far from the only possibility. Recently Pater went even further, claiming that the evidence in favour of a learning-based account is greater and more convincing than the evidence in favour of a rule-based account. Stating that:

> [i]n the absence of a specified theory of learning, however, the argument that a rich Universal Grammar (UG) of this type … is necessary to explain language acquisition is not completely solid. With the development of the rich theories of learning represented by modern neural networks, the learnability argument for a rich UG is particularly threatened. The question of how much and what kind of explicitly pre-specified linguistic structure is needed to explain language acquisition is in fact now receiving renewed attention in light of the learning capabilities of current neural networks. From a review of this work, it is hard to escape the conclusion that a successful theory of learning from realistic data will have a neural component. It is

> much less clear that a successful theory will need pre-specified UG parameters or constraints, though it seems likely that structured representations like those assumed in generative linguistics will play a role (Pater, 2018, p. 3).

In this case, Pater is not claiming that any rule-based account of language must necessarily fall away completely. Instead, he is changing the emphasis. Where Pinker and Prince claim that any successful connectionist network simply serves as evidence of a UG and the necessary rule-based account of language it is based upon, Pater reverses the formulation. He claims that any successful account of language acquisition must necessarily involve the neural account provided by connectionism, and that the best the rule-based account could hope for is structured representations arising from connectionist networks. However, the presence of structured representations would not mean that language acquisition reduces to the UG account, just that there is a generalised underlying structure to language. Some recent research in this area even points out that the neural and symbolist accounts may in fact be complimentary, with the neural account offering a potential solution to the grounding problem[18] faced by symbolists (Millière, 2024; Mollo & Millière, 2023).

When taking a step back from the details of these criticisms, what they amount to is something more interesting. For, when looking back from a temporal distance of more than 30 years, much of what they problematize is restricted to the limitations of connectionist networks of the 1980s and 1990s. In the span of years since they formulated their criticisms, connectionist networks, especially those networks working on natural language processing, have improved greatly. Seen in this light, the criticisms take on a different contextual milieu than originally intended. The above critics were educated in a paradigm dominated by classical computationalism, which made it easier to see the strengths of their own school, as well as the weaknesses of their rival. While this interpretation can help to make sense of the criticism, it also goes some way to explain connectionism's move towards the engineering paradigm. Faced with more knock-down arguments, similar in aim to that of Minsky's, the concern for any researcher invested in this field would be another AI winter.

So, besides the fact that engineering solutions greatly optimized connectionist networks, moving towards the engineering paradigm was also an excellent way to insulate the field from the criticisms of the classical camp. It meant that connectionism was no longer constrained by

[18] The grounding problem raises the issue of how symbols (such as words) connect to the real world objects and concepts to which they refer (Harnad, 1990).

bigger questions about whether these networks were capable of representation, or if modelling the brain would lead to the creation of strong AI. Instead, research in the field could be shielded from criticism by conceding that it was just "an invention for pattern recognition" (Rosenblatt, 1962, p. vi), which was precisely the kind of characterisation Rosenblatt argued was incorrect. Thus, while the concession increased network optimization, allowed connectionist research to secure lucrative funding from industry, and enabled research to continue uninterrupted, it also resulted in connectionism coming under the stewardship of a new paradigm.

1.3 The third epoch – engineering

The connectionist shift to the engineering paradigm does not just mean that the field is flooded with engineers, though, changes to research personnel remain a useful marker. The broader point here is that the shift is characterized by a change in methodological approach to building connectionist networks. Where early network research was focused on imitating brains, the third epoch marks a point of departure from this approach. Since, in the contemporary landscape "[m]any AI researchers don't care about how minds work: they seek technological efficiency, not scientific understanding. Even if their techniques originated in psychology, they now bear scant relation to it" (Boden, 2016, p. 7). This means that within this new paradigm, research on ANNs is undertaken without regard for how brains function, and without emphasis on explaining human behaviour. Where "[t]he PDP group proposed neural models mostly as new tools for exploring cognition, with a radical empiricist perspective of how the mind works … the deep learning research community is largely driven by application and market motivations, and indifferent to cognitive studies" (Perconti & Plebe, 2020, p. 10). This sentiment is perhaps best summed up in a 1996 paper by Warner and Misra. They explain that

> [n]eural networks have recently received a great deal of attention in many fields of study. The excitement stems from the fact that these networks are attempts to model the capabilities of the human brain. People are naturally attracted by attempts to create human-like machines, a Frankenstein obsession, if you will … From a statistical perspective neural networks are interesting because of their potential use in prediction and classification problems. Neural networks have been used for a wide variety of applications where statistical methods are traditionally employed. They have been used in classification problems such as identifying underwater sonar contacts and predicting heart problems in patients. They have also been used in such diverse areas as diagnosing hypertension, playing backgammon, and recognizing speech … As statisticians or users

> of statistics, we would normally solve these problems through classical statistical models ... [But it is] time to recognize neural networks as a potential tool for data analysis (Warner & Misra, 1996, p. 284).

This paper demonstrates the precise shift. Instead of seeing neural networks as an attempt to replicate the causal powers of the human, it became commonplace to see these networks as statistical tools - engineering their unique powers of pattern recognition towards the interests of industry and innovation. In other words, this new wave of researchers employs engineering methods to solve problems and optimize connectionist networks without taking direct inspiration from the brain (van Rooij et al., 2023). It does not matter whether the networks are optimized through mathematical techniques, ideas taken from physics, or statistical methods. All such techniques still represent a shift in methodology - away from imitating the brain and towards engineering ANNs to optimize functionality.

1.3.1 Shifting fields, personnel and methodology

Following from the success of the engineering methods employed by the PDP research group, one of the students of that group, Hinton, continued the work which they had begun. He would publish prolifically in the years which followed and continued to improve these networks using the engineering approach. For example, in 1995, he was the lead author on *The wake-sleep algorithm for unsupervised neural networks* (Hinton et al., 1995). Although an important paper, it did not change the connectionist landscape to the same extent as Rumelhart and McClelland's book, rather its contribution significantly refined the kind of networks which they had already pioneered. However, Hinton was not the sole author of this paper, he was now surrounded by mathematicians, physicists and computer scientists, such as Peter Dayan, Brendan Frey and Radford Neal. Dayan was a mathematician, Frey was a computer engineer and Neal was a computer scientist. Gone were the days where connectionist network research belonged to philosophers, psychologists, and logicians.

Hinton's paper is by no means an anomaly. Other important papers from this period, even ones without Hinton's name attached, had also been published by researchers of this ilk. Yoav Freund, sole author of *Boosting a Weak Learning Algorithm by Majority* (1995), was a computer scientist. *Replicator Neural Networks for Universal Optimal Source Coding* (1995) was authored by Robert Hecht-Nielsen, a mathematician. Nanda Kambhatla and Todd Leen, co-authors of *Dimension Reduction by Local Principal Component Analysis* (1997), were trained as a mathematician and physicist respectively, yet they were working on connectionist

networks. Even one of the most influential papers of the 1990s, *Gradient Based Learning Applied to Document Recognition* (1998), was by Yann LeCun, a computer scientist. Inspired by biological models of vision, LeCun found "Convolutional Neural Networks, that are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques" (LeCun et al., 1998, p. 1). The success of this research may have been limited to handwritten characters at the time, but the following decades would confirm LeCun's suspicions, that the principles underlying the success of CNNs would be broadly applicable to the field of visual processing. More detailed explanation and analysis of CNNs will be left until the next chapter. The point here, once again, is to demonstrate the change in methodology and personnel. The et al. in LeCun's paper stands in for the names of Leon Bottou, Patrick Haffner and Yoshua Bengio. Bottou was trained as an engineer and computer scientist, Haffner was trained as a data scientist, and the now famous Yoshua Bengio was similarly trained as an engineer and computer scientist.

The engineering approach to connectionism continued to improve these networks into the new millennia, and Hinton continued to be at the forefront of this research. In 2000, working on facial recognition, Hinton and Yee Whye Teh published *Rate-coded Restricted Boltzmann Machines for Face Recognition*. This paper is of particular interest because it appears to be between two of the connectionist epochs. Hinton and Teh were using methods from statistical physics for network optimization such as Boltzmann Machines[19] and Restricted Boltzmann Machines[20], while also admitting that the networks were still 'neurally inspired' (Teh & Hinton, 2000, p. 1). Notably, these networks are no longer being spoken about as brain models in the way that Rosenblatt did. Although this may seem a small shift, the move from being brain models to being just 'neurally inspired' is significant and only enabled by the shift to the engineering paradigm. The implicit assumption here is that the engineering practices applied

[19] A Boltzmann machine refers to artificial neural networks which employ a concept from statistical physics to pick up patterns in data. These networks train a hidden layer as a Boltzmann machine. A unique feature of these networks is that there is no traditional output layer, just an input layer and hidden layer. Additionally, every artificial neuron is connected to every other artificial neuron in the network. To begin, the weights start out stochastically, then, employing a technique called contrastive divergence, the hidden layer processes the data and compares the resulting functions against other resulting functions, repeatedly changing the weights until all functions are the same (or very close), which is when a Boltzmann machine can be said to have achieved a low energy state – which is its aim. In this way, a Boltzmann machine is trained in an unsupervised manner and learns through gradient descent. In other words, by constantly comparing all the different functions and adjusting weights to minimize the contrastive divergence between functions, a Boltzmann machine is able to find the function which best fits the given data (Hinton, 2007).

[20] A Restricted Boltzmann (RBM) machine differs from a Boltzmann machine in that the RBM's network of artificial neurons is not connected to all other artificial neurons in the network, rather the RBM is set up in a way that every artificial neuron is only connected to every artificial neuron in the next layer.

to these networks is moving them further away from brains. In other words, a basic connectionist network is an example of brain modelling, but a basic connectionist network plus the engineering approach is an example of something which is only 'neurally inspired'. Put simply, it appears that a new assumption has begun to take root within the engineering paradigm, that engineering a brain model may make it less like an actual brain.

Hinton would continue to be at the centre of this paradigmatic transformation of connectionism, working alone in *Training Products of Experts by Minimizing Contrastive Divergence* (2002) and with Teh to publish *Energy-Based Models for Sparse Overcomplete Representations* (2003). Fresh insight for this paper came from two new coauthors, Max Welling and Simon Osindero. Welling was a physicist, who's work in this paper, the application of physics concepts to connectionist architecture, would allow him to migrate into the field of connectionist AI. Osindero, following a similar trajectory, was trained in physics and mathematics but would also use this background to migrate into connectionist AI. Their contributions were especially useful because many of the engineering solutions which were driving connectionist research forward were mathematical techniques imported from statistical physics. And, in 2006, the transformation of connectionist networks into an engineering paradigm would bear some of its sweetest fruit. Hinton, Osindero and Teh would publish *A Fast Learning Algorithm for Deep Belief Nets* (2006). Then, later that year, the liminal paper founding the school of deep learning would be popularly disseminated in Hinton and Salakhutdinov's consolidation paper, *Reducing the Dimensionality of Data with Neural Networks* (2006).

1.3.2 Deep learning

The founding of the deep learning school of connectionist AI is particularly significant since it represented an advance as important as that of Rosenblatt, or Rumelhart and McClelland. Most of the research conducted after Rumelhart and McClelland's was within the paradigm their work had created and continued to only use one hidden layer. The assumption behind this practice was that adding more hidden layers would not improve the networks functionality. The following captures this claim:

> "[w]e have found no difference in the optimal performance of three and four-layered networks … [and that] four layer networks are more prone to the local minima problem during training … The above points lead us to conclude that there seems to be no reason

> to use four layer networks in preference to three layers nets in all but the most esoteric applications" (Perconti & Plebe, 2020, p. 4).

However, this assumption was ill founded, and Hinton's work would conclusively show that there were good reasons to use deep networks with more hidden layers. Once again, he would not discover it side-by-side with a psychologist, philosopher, or logician, but rather a computer scientist, Russ Salakhutdinov. The two of them

> succeeded in training a model with four hidden layers … by inventing a novel learning strategy, called the Deep Belief Network. The 'belief' was borrowed from the Belief Networks (Pearl, 1986), popular in expert systems, which Hinton appreciated (Perconti & Plebe, 2020, p. 4).

Though, the creation of deep learning networks was not quite so simple in practice. Pearl's network did not learn, so just adopting some of her research was not enough. Hinton also had to figure out that two layers of the network needed to be trained as Boltzmann Machines. Again, this is a clear example of the engineering shift taking place in connectionism. Boltzmann Machines were concept imported from statistical physics to optimize network functionality. Yet, this would be just the beginning of the transformation of connectionist research. Since, with the creation of deep learning networks, the engineering paradigm had well and truly begun. Deep learning would come to be known as "the most prominent and widely successful method in artificial intelligence" (Buckner, 2019, p. 1), and would continue to drive connectionist research over recent decades.

For example, another significant recent advance was *Generative Adversarial Nets* (GANs) (Goodfellow et al., 2014). GANs have inspired much of the generative connectionist artificial intelligence which has begun to characterize the landscape of the early 2020s. Familiar names show up here too with Welling and Bengio involved in Goodfellow's paper. However, like my mention of CNNs, more detailed analysis of GANs will have to wait until the next chapter. Though, the point remains fundamentally the same – this paper serves as yet another example of the shift firmly into the engineering paradigm.

1.3.3 Transformer networks

Finally, perhaps the most recent of the significant advances in connectionist networks is the Transformer. Transformers are neural networks which introduce 'attention' to ANNs and allow them to significantly increase training efficiency. They were pioneered in the paper *Attention*

is All you need (Vaswani et al., 2017). Though again, the details will be covered in greater depth within the next chapter. But, finding ourselves firmly rooted within this paradigm, a new problem is that:

> [the] commonly encountered attitude in these areas is that deep neural networks are just "more of the same" - perhaps an important engineering advance, but incremental rather than game changing - and so recent research developments do not merit the kind of careful scrutiny from philosophers that earlier waves of connectionism received … [Meaning that] philosophers have been largely silent on this technology so far (Buckner, 2019, pp. 1–2).

Though, considering the history in this chapter as well as the overwhelming successes of contemporary AI, evidence disputes the idea that these engineering advances are just more of the same (Millière, 2024). Especially since "deep learning neural networks have blown past predicted upper limits on artificial intelligence performance - recognizing complex objects in natural photographs and defeating world champions in strategy games as complex as Go and chess" (Buckner, 2019, p. 1). Rather, a more likely cause of the philosophical silence is the migration of connectionist research into the engineering paradigm. Although the engineering paradigm has insulated connectionism from philosophical criticisms and allowed remarkable progress to take place without concerns about what this implies for intelligence, it has only managed to achieve this by decoupling connectionism from its roots as a brain model. Therefore, the purpose of the next chapter is to investigate the grounds for decoupling connectionist networks from brain models, and to find what kind of similarities remain (if any at all). This will in turn enable us to justify the kinds of comparison which ought to be made about these networks, which is especially relevant since current comparisons range from the idea that none can be made at all, to the position that ANNs constitute some kind of mind.

2. Different Beetle, Same Box[21]

While the previous chapter demonstrated the origins of connectionist networks – specifically that the cradle of their creation was a brain model - this does not mean that modern iterations are brain models too. Moreover, since "[n]eural networks originally developed out of an interest in modeling the human brain ... [but have since] found applications in many different fields of study" (Warner & Misra, 1996, p. 292), the focus of this chapter will be investigating just how different and how similar brains and ANNs are.

2.1 Contemporary comparisons

Some argue that these networks are significantly different from one another and comparisons are unhelpful (Bender et al., 2021; Warner & Misra, 1996), while others object, arguing that a greater understanding of ANNs will help us to better understand the brain (Buckner, 2019; Linzen, 2019). In particular, the way connectionism was decoupled from its roots in brain modelling during the third epoch is cited as a reason to deny meaningful comparisons. The basic idea is that the engineering of these brain models has distanced connectionist networks enough from brains that meaningful comparisons are not justified. Put another way, previous epochs

> rightly recognised the tremendous potential of AI as a theoretical tool, but due to widespread, implicit makeist elements, AI and cognitive science became increasingly dissociated over time. Now, interest in AI among cognitive scientists is enjoying a renaissance - but the interest seems to be in the wrong type of AI, namely AI-as-engineering, which distorts our understanding of cognition and cognitive science.

[21] The idea for the title of this chapter came from Ludwig Wittgenstein's Logical Investigations. In §293 he uses the analogy of the beetle in the box to demonstrate the limitations of private language (Wittgenstein, 1986, p. 100). I have no intention of employing the analogy to say something about private language. Rather, for the purposes of this chapter the analogy is fruitful for a concise explanation of my general argument. Put simply, different beetle, same box refers to the idea that while engineering has changed connectionist networks, it has not changed them enough to make them incomparable to brains. This is due to the shared and numerous significant similarities across several different levels of analysis. This means that regardless of whether the beetle is smaller, larger, differently coloured, or even in the possession of a strange carapace, when we open the box and inspect the details, what we find is still a beetle.

> Accordingly, the time is apt to reclaim AI-as-theoretical-psychology as a rightful part of cognitive science (van Rooij et al., 2023, p. 14).

In this way, "it would be naïve to discount the potential contribution of engineering advances to scientific research" (Millière, 2024, p. 2) Accordingly, this chapter will investigate the grounds for comparison between brains and ANNs. I will make the argument that, upon closer inspection, there are several good reasons to justify such comparisons – especially when brains and ANNs are compared as information processing systems. In that regard, brains and ANNs have significant structural similarities, these structural similarities scale up to a similarity in how they transform information from input to output, which eventually scales up to computational similarities which take the form of high performance at tasks in the same domains. These similarities fit David Marr's Three Levels.

It is important to acknowledge at this point that such analogies are by no means a novel contribution. These comparisons have already been tentatively made by others about cognition, visual processing, language processing, etc. (Buckner, 2019; Linzen, 2019; Pater, 2018). For an example of one such comparison, consider connectionist networks and vision:

> if it is conceded that DCNNs[22] do explain human perceptual similarity judgments, there will remain significant debate in philosophy of science as to what kind of explanation they provide. Much of the enthusiasm for neural networks in the past has been derived from their supposed structural similarity to brains; a wave of 'new mechanism' cresting in philosophy of science … may seem well placed to vindicate this enthusiasm with a rigorous model of explanation … DCNNs, however, are pitched at a high level of abstraction from [the] perceptual cortex, which could lead one to doubt that they succeed at providing mechanistic explanations for even perceptual processing (Buckner, 2019, pp. 14 & 15).

In this case, it remains an open question whether ANNs process information in a manner similar enough to brains to provide a precise explanation of how humans process visual information. For Buckner, this means that while high level explanations of these networks remain underdetermined, this does not prevent mechanistic, lower-level explanations (Buckner, 2018, pp. 5369–5368). And, I am sympathetic to the claim that the significance of the changes wrought by the engineering paradigm makes any one-to-one comparison unlikely to be

[22] Deep Convolutional Neural Networks – a type of network specialised for visual processing tasks.

accurate - but brains and ANNs do not need to be identical to justify meaningful comparisons. Instead, they need to retain enough significant similarities to scale up to meaningful comparison regarding processing mechanisms, domain competencies, etc.

Visual processing is not the only area in which such comparisons are being made either. Linzen makes the claims that "deep learning and the scientific study of language can benefit each other" (Linzen, 2019, p. 106). This does not mean to say that brains and ANNs are processing linguistic information in an identical way. Rather, Linzen simply recognizes that that an understanding of one may well benefit an understanding of the other. From the one direction

> [l]inguists can contribute to research on neural networks for language technologies by clearly delineating the linguistic capabilities that can be expected of such systems, and by constructing controlled experimental paradigms that can determine whether those desiderata have been met. In the other direction, neural networks can benefit the scientific study of language by providing infrastructure for modeling human sentence processing and for evaluating the necessity of particular innate constraints on language acquisition (Linzen, 2019, p. 99).

The second point is of particular interest - that ANNs can contribute to our scientific knowledge of language by supplying infrastructure which can generate a model of human acquisition and use of language - perhaps even illuminating the partition between what is learned and innate. Here Linzen's notion that ANNs can supply such infrastructure indicates that there are enough significant similarities between brains and ANNs to infer from one about the other. Otherwise ANNs would not "provide a useful platform for constructing models of language acquisition" (Linzen, 2019, p. 106). The paper demonstrates this point with the example of Recurrent Neural Networks (RNNs) trained with structural biases which in turn outperform the standard model RNNs on some specific tasks (Linzen, 2019, p. 103). However, other scholars are willing to go further and argue that comparisons are justified on a claim, that "language acquisition proceeds primarily through data-driven learning of some form" (Clark & Lappin, 2013, p. 89), which suggests that we ourselves could be using something like an ANN to process linguistic information without the need for strong inductive biases[23]. Considering both positions, while the learning account has earned its place as seemingly essential to explanations of language

[23] Inductive biases refer to innate preferences or constraints which structure learning.

acquisition going forward, current indications do not provide enough evidence to completely rule out some component of innateness either.

Another instance of this type of comparison is present in the work of Pater - who is also supportive of such a move in linguistics, saying that "the continued general gulf[24] between these two lines of research is likely impeding progress in both: on learning in generative linguistics, and on the representation of language in neural modeling" (Pater, 2018, p. 1), which in turn makes a similar connection to the one Linzen made in the previous paragraph. Though, as discussed in chapter 1, Pater points out that current evidence provides strong indications that a neural component is essential for language acquisition, but that this is not necessarily the case for innateness claims like a UG. And his position towards language acquisition deserves some further consideration - especially if the learning component present in ANNs does give us a strong account of how language acquisition happens in the brain. Though, even if this account is not complete, and there is some innate language structure, what Pater is highlighting is that some level of neural processing is likely to remain essential to our account of language acquisition - which not only indicates similarity in how brains and ANNs process linguistic information, but also that humans may well be processing language in a more empiricist manner than previously thought. Nevertheless, the nature of language acquisition is still a hotly contested topic (Millière, 2024, p. 5).

Although the above examples are interesting, they do not form the basis for my argument. They simply show that there are already several researchers comparing brains and ANNs. Hence, in this context, I am not saying we ought to allow meaningful comparisons between brains and ANNs because others are doing it. Instead, my aim is to investigate where such intuitions are grounded and highlight the kind of comparison which can be justified based on current evidence.

2.2 David Marr and The Three Levels

While unravelling the hidden workings of vision, David Marr famously formulated what he called *The Three Levels*. These levels were intended to demonstrate "the different levels at which an information-processing device must be understood before one can be said to have understood it completely" (Marr, 1982, p. 24). I am by no means claiming this analysis will enable us to understand either system exhaustively - such framing is merely useful for

[24] In this context, the gulf Pater is referring to is the one between generative linguistics and ANNs.

structuring our analysis and preventing slippage into stronger claims about agency, intentionality, consciousness etc.

The first level is that of ***computational theory***. To focus our analysis at this level, Marr claims that we must ask: "[w]hat is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?" (Marr, 1982, p. 25). In other words, this level is concerned with the larger aim of the system. For example, a chess program would have the computational goal of winning chess matches.

The second level is that of ***representation and algorithm***. To focus our analysis appropriately at this level, Marr claims that we must ask: "[h]ow can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?" (Marr, 1982, p. 25). In simple terms, this level is concerned with the underlying principles which leverage the processing and transform input to output. Here it could be something like how the chess computer transforms information into chess moves with use of a search tree composed of branching if-then rules.

The third level is that of ***hardware implementation***. To focus our analysis here, Marr claims that we must ask: "[h]ow can the representation and algorithm be realized physically?" (Marr, 1982, p. 25). This level is concerned with the basic constitution of the system in terms of structure and physical hardware. For example, it can be a physical system like a Universal Turing Machine, which is the hardware that allows the realization of a chess playing computer. In the following pages, I will argue that brains and ANNs share similarities at all three levels, which in turn means that they are similar enough to generate meaningful comparisons of a particular kind, at least as information processing systems.

2.2.1 The level of hardware implementation

Marr organizes the Three Levels starting with the computational level and ending at the level of hardware implementation. My choice to do the opposite[25] aims to emphasize the idea that the realization of higher-level properties is contingent on the levels below, not vice-versa. In other words, it is the lowest level which forms the foundations from which the properties of levels above can begin to emerge.

[25] This is not to suggest that Marr was unaware of such considerations, it is simply a presentation difference.

To begin, at the level of hardware implementation, there are some obvious differences between brains and ANNs. For instance, the atomic composition of the two is distinct. One is made of carbon, water, etc. and the other is made of Silicon, Copper, etc. This is called substrate independence[26]. Yet, if we ask ourselves the question which Marr suggests we ought to, "[h]ow can the representation and algorithm be realized physically?" (Marr, 1982, p. 25), such realization of hardware is not necessarily contingent on atomic composition alone, it is also contingent on the especial arrangement of the components themselves and what the relationships between them represent. Marr intended that the level of hardware implementation would be, in the broad sense, about "the details of how the algorithm and representation are realized physically" (Marr, 1982, p. 25). Which is to point out that the structure of a given system is also an essential component of how a system is physically realised.

To illustrate this point, consider the principle involved in the process of siphoning a liquid. For a colourful example, let us suppose that we are petrol pirates, who, in the face of rising petrol prices have resorted to siphoning from unattended cars. If we wish to successfully siphon petrol, the most essential element is not necessarily whether our tube is metal, or plastic, or glass; it is the preservation of the structure of the system which makes fluid siphoning possible. This means that the tube itself must be continuous and without holes, the end of tube which is inside the petrol tank must be higher than the end inside the jerrycan, and suction must be applied to the jerrycan end of the tube to begin the process. It is this especial arrangement of hardware components which supports successful siphoning of petrol, not necessarily the atomic composition of the tube or liquid. For instance, if the tube were made of sugar, we would be able to siphon petrol with it, but not water. This is because sugar is water soluble and not petroleum soluble. Hence, the atomic composition only becomes an issue if the especial structure of hardware necessary for siphoning petrol breaks down, and the pieces cannot perform the functions which enable the system to work as intended. With this example in mind, consider how the hardware of brains and ANNs is implemented, and how the especial arrangement of hardware necessary for the realization of both is similar. At the level of hardware implementation, while each is realized with a different atomic composition,

[26] Substrate independence refers to properties which are realizable in more than one system of substrates. For instance, in the case of a mind "[w]e can consider a mind substrate-independent when its self-same functions that represent thinking processes can be implemented through the operations available in a number of different computational platforms. For example, if we can carry out the function of a mind both in a biological brain and in a brain that is composed of computer software or neuromorphic hardware (a hardware architecture with design principles based on biological neural systems), then that mind is substrate-independent. The mind continues to depend on a substrate to exist and to operate, of course, but there are substrate choices" (Koene, 2013, p. 146).

importantly, we still find structural similarity. Though, considering that connectionist networks have their origins in modelling the brain this structural similarity should come as no surprise. Still, it bears pointing out the precise ways in which brains and ANNs resemble one another structurally.

Firstly, both are composed of local units called neurons. In the case of the human brain, it "is made up of neurons as well as other cellular matter. The exact number of neurons in the human brain is unknown, but a recent estimate is 86 billion (Azevedo et al., 2009), with each neuron having tens of thousands of connections" (Ladyman & Wiesner, 2020, p. 57). Similarly, connectionist networks are composed of local units called artificial neurons; though, like brains, these are not the only elements which compose an ANN. However, unlike brains, artificial neurons are not composed of biological matter. Instead, they are often instantiated on a digital computer. Importantly, whether it is Rosenblatt's Perceptron, Rumelhart and McClelland's PDP network, or one of the newer deep networks, the local unit, the neuron, remains an essential part of the structure which brains and ANNs share.

Some may object that the name alone does not do enough to connect the neuron and artificial neuron – pointing out that we could have called one an apple and one an orange, but such a name does not make either of them fruit. This would be a fair point if all they had in common was the name. Yet, as discussed in the previous chapter, the history of early connectionist networks is a history of modelling the brain. This means that while the historical link may have inspired the name artificial neuron, it is not the only piece which was borrowed. For instance, our second similarity is that both the collection of cells we call biological neurons and the discrete local units we call artificial neurons function as information processors – serving as mathematical objects akin to multi-dimensional vectors which pass on information in a distributed fashion to the next layer in the network. This means that within brains and ANNs, both iterations of the local unit serve the same general function, passing on information from one layer to the next. Importantly, they also represent the same functional component within the structure of the especial system arrangement: a local unit responsible for receiving information from the previous layer and passing it on to the subsequent one.

Within the brain **"**the basic structure of all neurons is [also] that they have input and output regions. Information, in the form of electrical and chemical signals, flows from the former to the latter." (Ladyman & Wiesner, 2020, p. 59). And this is the same in ANNs – seeing as artificial neurons also pass information through the network via their input and output regions.

This is a common feature of connectionist networks which goes as far back as Rosenblatt's Perceptron. The 'neurons' in Rosenblatt's perceptron may not have been connected to one another, but they still passed information through the network, with the structure of the artificial neurons allowing the flow of data from input to output. Newer connectionist networks like ChatGPT employ the same principle, just on a much larger and more sophisticated scale. ChatGPT-3 had 175 billion parameters and ChatGPT-4 over 100 trillion parameters (Leib, 2023, p. 425). From a technical standpoint, the fact that artificial neurons also function as information processors is evident insofar as they also adhere to

> three simple sets of rules: multiplication, summation and activation. At the entrance … the inputs are weighted … [and] every input value is multiplied with individual weight. In the middle section … [there is a] sum function that sums all weighted inputs and bias. At the exit … the sum of previously weighted inputs and bias is … [the] activation function (Krenker et al., 2011, p. 3).

Therefore, despite rapid growth in the size of these networks, the local units still largely serve the same general function, taking in a wide variety of fragmented information, amalgamating it during their processing and then reducing it to an output like an activation function to pass it on. Hence, like biological neurons, artificial neurons also function as information processors.

The third similarity in structure is that which links the local units and allows them to pass on information from one layer to the next: both brains and ANNs have distributed connections between the local units in their respective networks. As mentioned previously, an essential element of the brain is "each neuron having tens of thousands of connections" (Ladyman & Wiesner, 2020, p. 57). This structural element of brain networks is present in ANNs – since artificial neurons also possess large numbers of interconnections with other artificial neurons within the network. Ironically, while Rosenblatt's Perceptron, had some of the strongest claims about its similarity to brains, it was the least like a brain in this regard. However, this is simply a consequence of its age, since it is one of the oldest and most rudimentary connectionist networks. Yet, such a feedforward network lacking in direct connections between neurons is very rare today. PDP networks, CNNs, RNNs, GANs, Transformers, etc. all have interconnected neurons (Buckner, 2019, p. 5). Thus, we find a third structural similarity in terms of how these systems are realized physically. Both make use of interconnections between local units to pass information into the next layer of the network. While it is important that information is passed on to the next layer of the network, the variation in ANN networks and

network tasks means that many networks do not necessarily connect every single artificial neuron in every layer to every other artificial neuron in the layers before and after. Instead, in some networks like DCNNs, an important feature is:

> [the] sparse connectivity between layers. Whereas nodes in Golden Age networks were often fully connected to each node in the layers above and below, nodes in DCNNs are usually only locally connected to nodes with spatially or temporally overlapping input receptivities … This sparse connectivity produces further gains in efficiency, because it greatly reduces the number of parameters that need to be learned, relative to a fully-connected network with the same number of nodes; it also makes computation more efficient … because the activation functions which need to be computed have far fewer inputs (Buckner, 2019, p. 7).

This illustrates the point that while different networks may arrange the connections differently, the common structural feature which all these networks share, is the possession of distributed connections between the local units themselves. Since, even if the particulars of how they are connected may reasonably vary, and some parts of the network may remain isolated from other parts, all connections still function as a link between artificial neurons; thereby allowing the networks to pass on information.

Taking a closer look at the connections themselves, we find a fourth and fifth similarity, since not all connections between neurons are connected equally. Specifically, in both brains and ANNs, connections between neurons are weighted and these weightings can differ. In brains:

> [i]nterconnections between neurons, in the form of synapses, are developed and maintained by feedback. By firing, a neuron will excite or inhibit other neurons, and the effects of those excitations on other neurons can reach it on a time scale comparable to that of its own dynamics. Repeated firing strengthens the semi-permanent synaptic connections, establishing memory and learned behaviour, which are associated with groups of neurons that fire together in roughly the same pattern each time they are activated (Ladyman & Wiesner, 2020, p. 59).

This points out that connections between neurons in the brain are weighted, meaning that some connections will be stronger and others weaker. This entails that some will require less activation to reach their excitation threshold and pass on information, while others will require more and will be inhibited and not pass on information. In brains these weightings are established by a variety of factors such as how regularly pathways are used or even as the result

of species-specific biases. In ANNs there can be far more bespoke control of individual weighting and individual weights can be changed manually; or the weighting of a network can be changed generally via techniques like gradient descent[27]. However, the point remains – in both networks weighted connections between neurons are a crucial component of the implementation of the structure of the network and in both cases the weighting determines whether the information will be passed on by the neuron (excitation) or not (inhibition). This similarity applies broadly to both brains and ANNs but not all networks rely on binary output possibilities like excitation or inhibition. For example, in a network more specialised for the task of processing visual information, such as a DCNN, the output of the artificial neurons can be scalar as opposed to binary. Buckner explains this point as:

> The most popular downsampling function in state-of-the-art DCNNs is max pooling, which sends up activation only from its most highly-activated input … In other words, if a max pooling unit receives input from a vertical line kernel and a horizontal line kernel at a particular location, then it will only pass activation to the next layer for whichever of the two was most strongly activated (i.e., it forces the layer to make a decision (Buckner, 2019, pp. 4–7).

Max pooling is an example of scalar activation in local units. To distinguish the more activated local units from less activated ones, the units themselves cannot simply have a binary output of excitation or inhibition. Instead, they possess a scalar activation which indicates not whether they are activated, but rather the degree to which they are activated. For example, instead of a binary output like 0 or 1 representing activation or inhibition, it would be a scalar output between 0 and 1 representing the strength of activity, like 0.37. Regardless of these variations, the larger structural elements in question remain - both that the connections between local units are weighted, and that these weightings can differ from local unit to local unit.

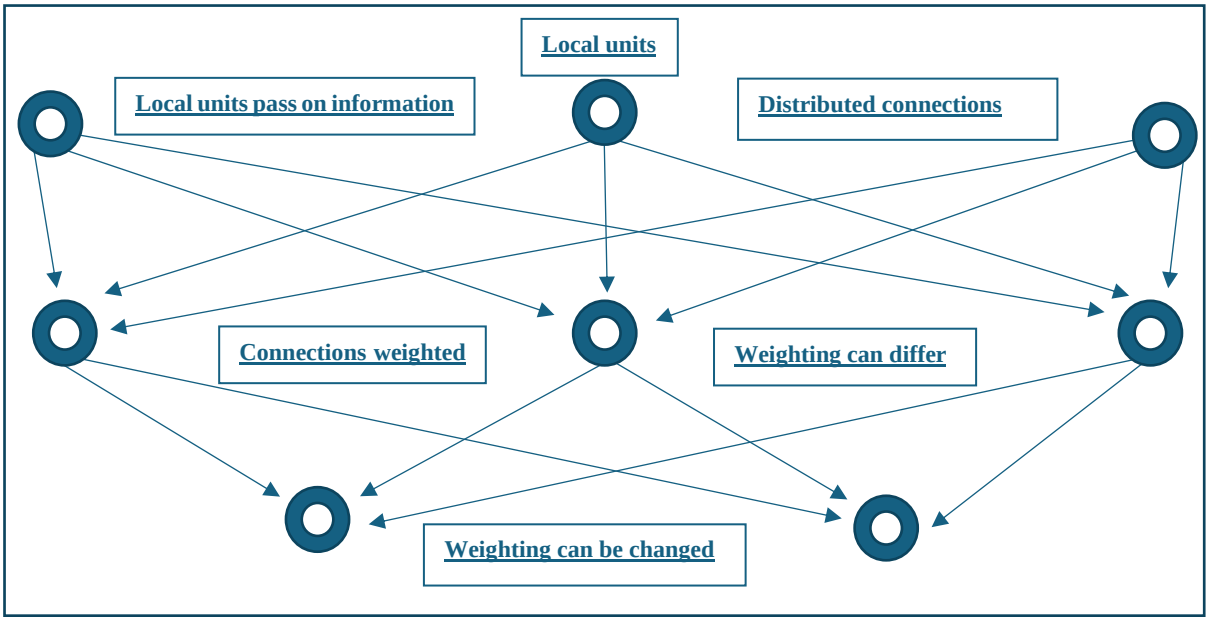
The sixth and final similarity at the level of hardware implementation is that the weights between connections in both brains and ANNs can change. The brain is able to change the weighting between the connections - since "repeated firing strengthens the semi-permanent synaptic connections" (Ladyman & Wiesner, 2020, p. 59). The more that the connections are used, the stronger they become. This results in a change in weighting which reduces the

[27] Gradient descent refers to a variety of mathematical techniques aimed at reducing the difference between the outputs the network generates and the true labels for those outputs in the training data. Gradient descent can be done using batch methods, stochastic methods, and methods which involve a combination of both (LeCun et al., 1998, p. 2319).

threshold required for a neuron to become active. The same applies to connections which are used less often. The absence of use changes the weighting and increases the threshold required for excitation of a neuron. Comparing this with ANNs, the specific way we can change the weighting of connections may well be distinct from brains (insofar as we have more control over these changes), but the fact that weighted connections can change remains common to both brains and ANNs. This similarity is not limited to one kind of ANN either. While in Rosenblatt's day the weights were changed manually, or in Rumelhart and McClelland's day networks were capable of automatically adjusting their own weights, or today ANNs employ Boltzmann machines and other sophisticated techniques for gradient descent; what is common to all ANNs is that they change the weighting of connections to optimize the function. Hence, even if the techniques employed may vary across ANNs, like brains, ANNs too share an ability to change the weighting of their connections.

Importantly, this ability to change the weighting generates some of the structural flexibility characteristic of these learning algorithms in both brains and ANNs. For, the kind of activation patterns which will come to represent the transformation of information from input to output require, precisely, this kind of structural flexibility to map a variety of relations of statistical dependence, spatially (Cain, 2016; P. M. Churchland, 2012; Elman, 1990). The variety of similarities at the level of hardware implementation can be seen in figure 4.

Figure 4: *Structural similarities between brains & ANNs.*



2.2.2 The level of representation and algorithm

Building on the foundations established at the level of hardware implementation, we must look at the next level and ask "[h]ow can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?" (Marr, 1982, p. 25). We find that, while some of the details do differ, the broad principle by which information is transformed in brains and ANNs remains largely the same. The processing of both is reliant on high dimensional space to map relations of statistical dependency.

To illustrate this, consider the kind of processing which such hardware enables. A network of local units possessing the structural similarities discussed at the previous level necessitates that information is processed in a distributed fashion. Local units process fractions of information and pass them on to the units in the subsequent layer, which in turn does the same. In brains this means that the processing of the network does not take place in any single neuron, instead the load is distributed throughout the neural network. This in turn leads to phenomena like "distributed decision making and a flexible division of labour." (Ladyman & Wiesner, 2020, p. 61). Of course, processing in an ant-colony or a LAN network also takes place in a distributed fashion, so this alone does not mean that ANNs process information like brains do. Though, in ANNs, much the same process is occurring and has been standard practice since parallel distributed processing was pioneered in the late 20th century (Rumelhart & McClelland, 1986). Across a variety of networks, each artificial neuron is connected to other artificial neurons in the prior and subsequent layers[28] – which means that what went into the network as a whole piece of information is distributed across the network and then processed in parallel. While such similarities are interesting, they are still too broad to demonstrate any meaningful comparison at this level. Instead, what we require is a similarity in "the choice of representation for the input and output and the algorithm to be used to transform one into the other" (Marr, 1982, pp. 24 & 25). In other words, what we need to show is that the representation which brains and ANNs use to transform input to output relies on the same underlying principle. Accordingly, the fact that both represent transformations using activation patterns is of particular interest. Though, activation patterns only tell us what represents the transformation, not the underlying principle behind it.

[28] Apart from the input layer and output layer.

For example, in ANNs, researchers working in explainable AI employ activation patterns to give an account of how information is transformed from an input to output. Below, the activation patterns function in a way which allow the network to identify a picture of a butterfly:

> [b]y extracting similar activation profiles within the high-dimensional activation space of a neural network layer, we find groups of inputs that are treated similarly. These input groups represent neural activation patterns (NAPs) and can be used to visualize and interpret learned layer concepts (Bauerle et al., 2022, p. 1).

Of course, 'learning' may be a generous interpretation of what is happening within a neural network being engineered to identify images of butterflies, but the point stands. The 'neural' activation patterns in ANNs are a representation of the kinds of patterns which transform the input from mere pictures of animals to an output function which can distinguish butterflies from other species. This manner of representing the transformation from input to output with activation patterns is shared with brains.

In brains, the presence of activation patterns is identifiable beyond a mere anatomical network - since such patterns do not need to be represented by a direct physical connection. Instead, activation patterns can be represented by functional networks. These are things like "patterns of statistical dependence among neural elements … For example, if activity in one brain region occurs when some other brain region is active, and vice versa, there is a functional connection between those two regions" (Yan & Hricko, 2017, p. 2). This is an expansion of the kind of activation pattern discussed earlier in the context of ANNs, but at its core it is the same principle by which the transformation of information from input to output is represented. The difference with functional networks is that there is no need for an actual, physical connection between one region of the brain and another; what is essential is simultaneous activation which in turn indicates a different kind of network, a functional one. Relevant to the point is that the processing of such a functional network in the brain is represented by activation patterns (Yan & Hricko, 2017, p. 2). This is even more relevant when we consider that ANNs have also been understood as functional networks for quite some time already, and that such an approach has been highly useful for our attempts at interpretability (Castillo et al., 1999).

However, as mentioned, activation patterns are an explanation of what represents the transformation of input to output within brains and ANNs. This similarity may indicate the presence of a shared underlying processing principle, but it is not the principle itself. It explains that both brains and ANNs represent the transformation of input to output with use of a similar

strategy, neural activation patterns, not why this is the case. To provide an explanation of the underlying principle behind neural activation patterns, we must understand the kind of task the pattern is executing within the context of the larger network. Here, the work of Elman from the first chapter is of use once more. Within the context of ANNs and language:

> activation patterns pick out points in a high (but fixed) dimensional space. This is the space available to the network for its internal representations. The network structures that space in such a way that important relations between entities is translated into spatial relationships. Entities which are nouns are located in one region of space and verbs in another (Elman, 1990, pp. 205–207).

Considered this way, an activation pattern is a spatial technique for representing complex relationships between variables. Together the local units and the connections between them form the space the network has available for its representations; the activation patterns then are the specific representations of given statistical relationships. Such relationships of statistical dependency just take on a spatially represented form within the network. The relations themselves are not spatial, they are simply spatial abstractions of real relationships within a given set of information which the high dimensional space of the network is able to map in the form of an activation pattern. Hence, the type of variable discussed by Elman here, language, is not the only kind of variable these networks are capable of mapping. If one is to give these networks visual information, like the butterfly example from Bauerle, the network will use high dimensional space to map relationships between the visual variables and find the activation pattern (or groups of activation patterns) which represent a butterfly. In the case of a butterfly, instead of nouns, the activation pattern is mapping shapes and the underlying principle by which it does so remains the same. It is using activation patterns within the high dimensional space created by a network of local units to represent relationships of statistical dependency spatially. Perhaps one such relationship is the angle of the butterfly wings (or a visual variable even more fractional than that), but across different domains, high dimensional space is being used to map the relationships between variables, and it is the activation patterns in this space which come to be the basic units which represent the transformation of information from input to output.

Such work would later be built upon, and defended, by Churchland (P. Churchland, 1996a, 1996b, 1998), eventually culminating in his book *Plato's Camera* (2012). Although he employs this account to make claims about higher level properties like mental content (Prinz, 2005) - a

claim which I am not addressing in this paper - his analysis of how these networks process information spatially adds a lot of detail to the work Elman began. For instance, Cain gives a condensed version of Churchland's analysis, explaining:

> patterns of activation here have crucial representational significance and, indeed, serve to encode concepts … [and] that distinct possible patterns of activation at a particular layer of a network stand in distance relations to one another, with some patterns being closer to one another than others … In short, then, activation patterns in a given layer are points in a multidimensional space where each point bears a distance relation to all other points (Cain, 2016, pp. 35–36).

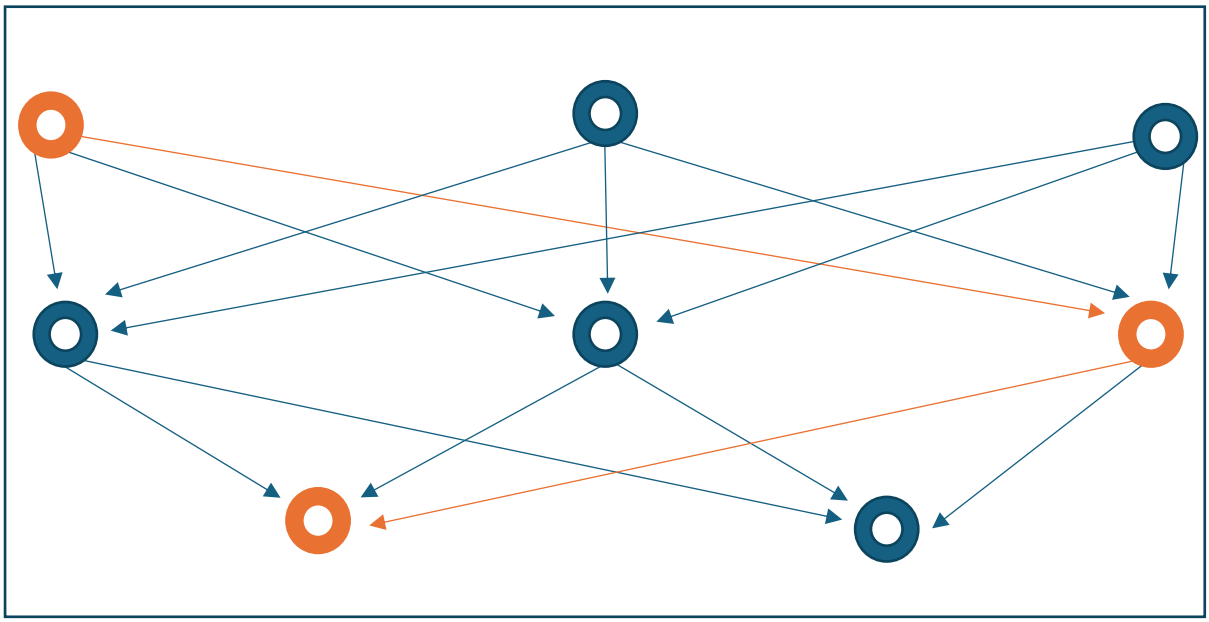
Perhaps we can challenge some of the details of the Churchland view, such as the claim that it is distance, specifically, which encodes the information fragments of what will come to form an activation pattern for a larger concept. Yet, the broader point here will still allow us to build on Elman's account. Since, even if distance is an oversimplification of the weighting of a brain or an ANN, the idea still captures the important relationships between parts and whole which enable this kind of processing. Regardless of the details of the weighting, what such a connection between the local units captures is a statistical relationship in the information being mapped spatially. Though, the above is referring specifically to ANNs, not brains. Relevant to brains is the example of visual processing in primates. Here:

> the activity in each distinct cortical area is itself retinotopically organized, so that the arrangement of neurons on the cortical surface mirrors the arrangement of their receptive fields (i.e. the parts of visual space that they are responsive to). In this way, neurons responsive to nearby regions of space are also located close to each other on the cortical surface. Together, the receptive fields of the neurons in each brain region completely cover the visual field … [Which] thus contains a spatial map of the whole visual field, composed of an array of these locally responding regions (Cao & Yamins, 2021, pp. 10 & 11).

The way brains utilize the high dimensional space available to them to represent complex relations between visual variables is especially obvious in the example above. The closest neurons literally reflect the arrangement of the visual fields which they are receptive to, mapping them completely by using high dimensional space for representations. Though, this kind of visual processing is not limited to primates. It extends to both the brains of other animals and to other ANNs (Buckner, 2019, pp. 8 & 9). Hence, at the level of representation and

algorithm, the underlying principle which leverages the representation of the transformation from input to output is preserved between brains and ANNs. Both use high dimensional space to represent complex relationships between variables by mapping the statistical dependencies of these relationships spatially. Moreover, the larger representation of relations of statistical dependency take the form of neural activation patterns in both brains and ANNs, storing information about these relationships through specific combinations of patterns in high dimensional space.

Figure 5: *An example of an activation pattern (orange).*



Here it is important to clarify the kind of claim being made. I am not saying that ANNs model the human mind, nor does the evidence support such a powerful claim. Instead, I am pointing out that brains and ANNs transform information with use of the same underlying processing principles – which supports a weaker claim for comparison – such as ANNs are a mechanistic abstraction of neural processes (Blank, 2023). Though, even this claim is larger than the one I am making, and I am actually unwilling to claim ANNs are even a mechanistic abstraction of neural processes in general. To be more precise, my claim is an even weaker one - that ANNs are a mechanistic abstraction of the spatial mapping processing principles present in neural networks. Considering the continued gulf between what ANNs and brains can do (Chaves & Richter, 2021), this smaller claim corresponds appropriately with ANNs current capabilities, not potential future ones. Moreover, the weaker claim still gives a satisfying account of even the most remarkable developments in ANNs. For instance, there has been some success with

ANNs being able to ‘read’ what a person is thinking from MRI scan information (Osborne, 2023). Such remarkable advances make it easier to argue for the stronger claim, but the weaker claim explains such a phenomenon too. This experiment was successful in using ANNs to transform activation patterns from MRI mappings of biological brains into surprisingly accurate interpretations of verbal cues (Osborne, 2023). Though, if ANNs are a mechanistic abstraction of spatial mapping processes in brains, these kinds of successes are to be expected. For, the ANNs can represent the same patterns of statistical dependency using the same underlying principle to transform the input to output. In fact, it would be more surprising if they possessed this mechanistic abstraction of neural processes and came to radically different answers. In this sense, the heart of the weaker claim is simple, it is the idea that "the number of neurons and the way they are connected is crucial for the cognitive capabilities of any organism" (Ladyman & Wiesner, 2020, p. 57). As has been shown at the level of hardware implementation, the basic structure of neurons and connections between neurons is preserved in the instantiation of ANNs. Accordingly, if the basic structure responsible for the capabilities of brains is present in ANNs, and that this structure in ANNs allows for a mechanistic abstraction of the brains spatial processing, it should come as no surprise that these networks can perform well at the same kinds of tasks that brains perform well at. This brings us to the third level – computational theory, or the (not so) coincidental presence of similar domains of functional competency.

2.2.3 The level of computational theory

At this level, the question we must ask is, "[w]hat is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?" (Marr, 1982, p. 25). In this regard, brains and ANNs are aimed at many of the same computational goals; and importantly, both demonstrate high levels of functional competency in these shared areas (Millière, 2024). Though, a caveat must be made here. In terms of the goal of the computation, neither brains nor ANNs are defined by a singular aim. Instead, they perform well at a variety of similar tasks. These shared domains of functional competency are most evident in the processing of visual information (Cao & Yamins, 2021), auditory information (Osborne, 2023), natural language (Chaves & Richter, 2021; Elman, 1990; Linzen, 2019), as well as the generation of art (Cetinic & She, 2022; Goodfellow et al., 2014). Hence, our analysis shall begin with these shared domains of functional competency.

In terms of visual processing, brains and ANNs both demonstrate high levels of competency; and ANNs such as CNNs, and the subsequent DCNNs perform especially well in this domain. As far back as 1998, the original CNNs were "shown to outperform all other techniques … provid[ing] record accuracy on business and personal checks … [Even being] deployed commercially and read[ing] several million checks per day" (LeCun et al., 1998, p. 1). At that time, the functionality of CNN visual processing was narrow, because its successes were largely limited to the recognition of handwritten characters. This is no longer the case. Today their powers of feature extraction have made their applications incredibly broad in the field of visual processing, and they have expanded from handwritten characters to animal classification (Bauerle et al., 2022), object classification, as well as facial recognition and beyond (Buckner, 2019). All of which are traditional domains of functional competency for brains. Yet, such proficiency does not provide enough evidence to claim that they process visual information in the same manner, just that visual processing is a shared domain of functional competency. It could of course be coincidence that they perform well at the same visual tasks. After all, CNNs and DCNNs are engineered specifically for this purpose.

The high performance of ANNs in the field of visual processing is not the only domain of functional competency – another is language. Consider the successes we have had with Large Language Models (LLMs) in recent years, and that they are as competent or occasionally even more competent than humans at processing language (Bender et al., 2021; Linzen, 2019). Perhaps the most practical example to demonstrate just how competent LLMs have become is that they now pose a threat to jobs in the copywriting industry (Karakas, 2023; Wu, 2024). Recent research

> conducted a large-scale public online Turing test with human participants and GPT-4. One GPT4 witness, Dragon, deceived users into believing that it was human fairly robustly across 855 games. As far as we are aware this is the first empirical demonstration of an agent achieving a 50% success rate at the Turing test on such a large sample (Jones & Bergen, 2023, p. 9).

Such significant developments have, particularly in the case of ChatGPT, brought the AI conversation back into the public eye. Yet, despite rapid advances, the Transformer networks responsible for LLM successes still suffer from processing issues which humans do not (Chaves & Richter, 2021; Linzen, 2019). Consequently, despite optimism that these networks will yield insight into human language use and acquisition, it remains an open question if or

when this will happen. Importantly, such questions rest upon whether these networks handle information in a manner similar enough that our understanding of one would aid our understanding of the other. However, regardless of how similar they are in that regard, the success of these networks in the domain of natural language processing still provides another domain of functional competency shared by brains and ANNs - language processing.

Visual and language processing are not the only domains of function competency either. Some ANNs are showing promise in human domains such as art, music, speech, video, etc. In particular, AI art has developed rapidly in recent years. The generation of AI art can involve a variety of different ANNs. Early successes in generating AI art, music, etc. (Goodfellow et al., 2014), were pioneered with use Generative Adversarial Networks (GANs), and these networks

> represent a turning point in the attempt to use machines for generating novel visual content. The key mechanism of a GAN is to train two "competing" models … implemented as neural networks: a generator and a discriminator. The goal of the generator is to capture the distribution of true examples of the input sample and generate realistic images, whereas the discriminator is trained to classify generated images as fake and the real images from the original sample as real (Cetinic & She, 2022, p. 9).

CNNs have also been involved in the development of AI art. In particular, the feature extraction abilities of CNNs enable sophisticated classification of visual information which has become an invaluable component to creating such art[29]. For example, CNNs have allowed for rudimentary distinction between 'content' and 'style' within the sample sets of artwork, as well as detailed automatic breakdowns of the constituent elements contained therein (Cetinic & She, 2022, p. 8). With that being said, the most recent networks involved in AI art are the kind which turn text prompts into artistic renderings, and these multimodal applications are largely leveraged by the adoption of Transformer networks (Cetinic & She, 2022, p. 10). These breakthroughs in the generation of AI art have resulted in a level of competency which is creating serious competition for human artists. In fact, the "impact of image generators on the art community, rang[es] from economic loss, to reputational damage and stereotyping" (Jiang et al., 2023, p. 372). Considering that AI image generators are proficient enough to pose a cultural and economic threat to human artists, we do not have to comment on whether such

[29] There is a burgeoning debate on whether AI content generated from human artworks can be considered art. As interesting as that line of inquiry may be, it is not relevant to this research and will not be dealt with here.

creations represent the same process as an artists' to point out that they are an example of a shared domain of functional competency.

Though, the presence of these shared domains of functional competency constitutes only a component of the computational level – it reveals that brains and ANNs are being aimed at many of the same kinds of goals. The next step is to investigate the logic of strategy employed by brains and ANNs to achieve these goals, and the appropriateness of such a strategy. In that regard, Marr's second level will be of further use. At the second level we began a discussion, not just about how transformation of input to output is represented, but also about the underlying principles which make these representations successful. The key takeaway was that brains and ANNs both process information by using high dimensional space to map relations of statistical dependency. By taking advantage of high dimensional space to map these relationships, what brains and ANNs are doing is still just pattern recognition. Consider it this way, obviously there remain significant differences between networks like DCNNs, Transformers, GANs, etc., and of course many of them are even aimed at different goals, but the logic of the strategy by which the computational goal is carried is still largely the same. Like brains, all the varieties of ANNs discussed also rely on the same kind of spatial processing strategy to recognize patterns.

For instance, within the first domain of functional competency, visual processing, "when limiting deep learning to convolutional models for vision, there is growing evidence of striking analogies between patterns in these models, and patterns of voxels in the brain visual system." (Perconti & Plebe, 2020, p. 7). If we think back to the details about visual processing in brains and ANNs, the logic behind the computation remains a spatial one for processing information. Which is to say, the "nodes in DCNNs are usually only locally connected to nodes with spatially or temporally overlapping input receptivities." (Buckner, 2019, p. 7). This is a strategy also employed by brains to extract visual features. In fact, the original Convolutional Neural Networks (CNNs) were inspired by the work of Hubel and Wiesel on visual processing in cat brains (1967), though further research in neuroscience has shown that this is broadly representative of mammalian visual processing too (Buckner, 2019). Hence, the shared retinotopic organization of local units which spatially mirrors visual receptive fields in brains and ANNs goes a way towards explaining the successes of DCNNs in this domain (Cao & Yamins, 2021, pp. 10 & 11). Importantly, while the question of how much, precisely, the processing of one resembles the other remains an open question, what is clear is that spatial processing is present in both. Therefore, networks like CNNs are an example - not only of a

computation with a similar goal to brains - but one where the logic of the strategy by which the computation is carried out to achieve these goals is largely the same too – both use high dimensional space to map relationships of statistical dependency and identify real visual patterns.

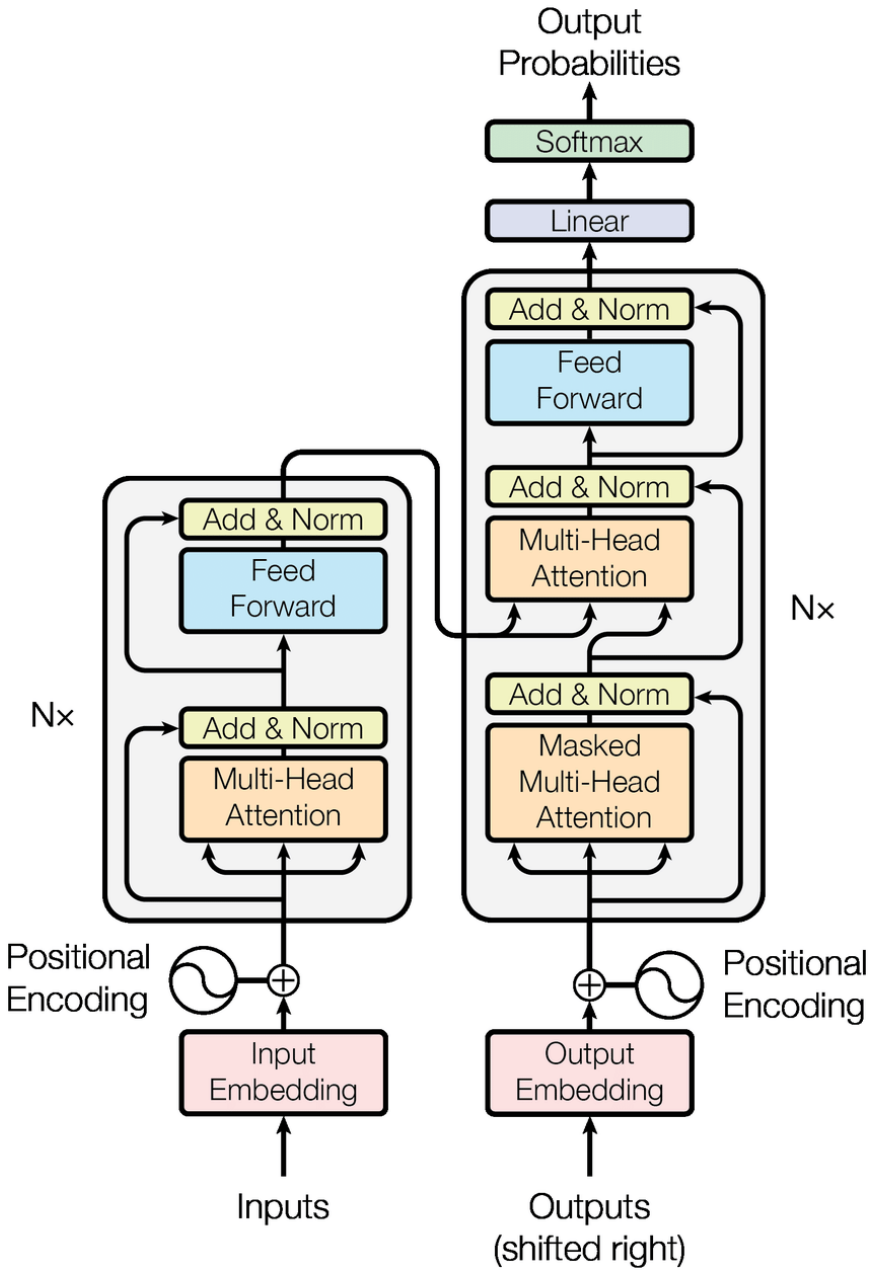
While CNNs may not be utilized much in the domain of language, the logic of the strategy by which the computation is carried out by the ANNs relevant to the language domain is largely the same. For instance, earlier networks aimed at language processing, like RNNs, employed high dimensional space in order to represent different types of words, like nouns, verbs, adjectives, etc. as well as distinguishing more sophisticated parts of language like token usage (Elman, 1990, pp. 205–207). While RNNs are no longer the preferred architecture for language processing tasks today, the logic of the strategy employed by newer networks remains largely unchanged too. These newer networks, Transformers, have become incredibly popular due to their accuracy with sequential information. Transformers themselves did not introduce much new individual technology to the field of ANNs, instead they combined a variety of existing techniques into a uniquely effective architecture. In simple terms, Transformer networks have an architecture made of two larger internal structures, an encoder[30] and a decoder[31](Vaswani et al., 2017). The encoder handles the input, and the decoder handles the output. The encoder

[30] The encoder handles the input and is constituted by two basic layers, "[t]he first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network" (Vaswani et al., 2017, pp. 2–3). Before the information passes to the first layer, each piece of sequential information, say a word, is positionally encoded by giving it a number which represents its position in the information. Then, in the first layer, the self-attention mechanism processes each of the positionally encoded pieces of information with use of a multi-head system of attention which can individually process and assign weights to each of the positional encodings. This allows the network to pay basic 'attention' to important pieces of information. The second layer then involves feeding this information through a neural network to map the relations of statistical dependency. If the Transformer network is aimed at language, like ChatGPT, the self-attention mechanism is assigning weights to words in a given input question and then running the question through a feedforward neural network trained on language. Even this is an oversimplification though, since these two layers are repeated six times in the encoder and subject to normalization after each layer. This entire structure is what is called an encoder (Vaswani et al., 2017, pp. 2–3).

[31] It is similar to the encoder, still engaging in positional encoding to start with, except its first layer now involves masked multi-head attention and it "inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack" (Vaswani et al., 2017, p. 3). Hence, the larger structure of the decoder is as follows. It begins with a layer of masked attention - which allows the attention units to see only the previous word in the string. This is performed in the decoder and not the encoder because the output generated by the decoder needs to make sense sequentially. Next, the output that the encoder generated is inserted into the next layer along with the output from the masked layer. This layer is not masked. Finally, in the third layer the output from the previous layer is processed by a feed forward neural network. Again, this is an oversimplification because these three layers are repeated six times and subject to normalization after each layer. This entire structure is what is called a decoder (Vaswani et al., 2017, p. 3).

uses a self-attention mechanism to derive important context about the input and attend to it appropriately. Below is the original structure of a Transformer.

Figure 6: *The structure of a Transformer (Vaswani et al., 2017, p. 3).*



This (positional encoding) was innovative and quite a revolution in the field, allowing the encoder to effectively process the emphasis of semantically and syntactically complex questions and requests. However, the information the encoder 'learns' about the input is not isolated from the decoder. The output function of the encoder is built into the very structure of the decoder. Hence, the decoder can apply the derived context 'learned' by the encoder to focus its 'attention' on relevant information while still generating a sequentially sensible string

of language. Yet, in the case of both the encoder and decoder, a feed forward neural network remains essential to the logic of the strategy by which it achieves its computational goal (Vaswani et al., 2017, pp. 2–3). The attention components merely enable Transformer networks to direct their attention to relevant information. Yet, the information itself is still being mapped spatially by the feed forward neural networks present in both the encoder and decoder, self-attention mechanisms just allow the network to allocate processing resources more efficiently and take better advantage of the relations of statistical dependency identified by the feed forward neural network reasoning. Therefore, in language processing, not only are Transformer neural networks an example of a shared domain of functional competency insofar as the computational goals they are aimed at are the same, but much like brains, Transformers rely on high dimensional space to map relationships of statistical dependency. Hence the logic of the strategy by which the computation is carried out to achieve these goals is largely the same too.

Within the domain of AI art, the variety of networks involved in the generation of AI art does pose some difficulty for analysis. The three main ANNs involved in the generation of AI art are CNNs, Transformers and GANs. As has already been discussed, networks like CNNs use high dimensional space to map relationships between complex visual variables, much like brains do. In fact, as was mentioned earlier, CNNs are more literal in their representations since they actually mirror the visual field itself and are directly inspired by work done on the processing of visual information in cat brains (Buckner, 2019, pp. 8–9). Hence, the similarity in the logic of the strategy used by CNNs for their computations has already been established as spatial reasoning. Additionally, Transformer networks were covered in our analysis of linguistic ANNs. They do have complex self-attention mechanisms which distinguish them from other ANNs, but the self-attention mechanisms direct their attention to relevant information which has already been mapped using the spatial reasoning of a feed forward neural network. As a result, the logic of the strategy by which they achieve their computations is still leveraged by the same kind of spatial reasoning. Finally, this leaves GANs to cover.

GANs can involve a variety of different ANNs depending on the data being optimized by the generator and discriminator. The type of generators and discriminators can also vary based on the type of information being worked with. What is unique about GANs is that they enable two ANNs to compete with one another to better optimize the function. This competitive element to the training of GANs does distinguish them from other networks, but it does not alter the fundamental logic of the strategy by which the computation is carried out. While there may be

variation among different networks - which makes them more appropriate for some tasks and less appropriate for others - the base structure of GANs still relies on neural network architecture, and these networks rely on the underlying processing principle of high dimensional space to map relations of statistical dependency. Simply put, GANs still make use of high dimensional space to map these relationships, they just also employ additional networks to 'mark' the output as well.

2.3 What kind of comparison can be made?

Considering these similarities, the claim that ANN engineering has distanced them enough from brains to prevent meaningful comparison has grounds to be challenged. While it is fair to point out that "[t]he progress of deep learning over the past decade has been more significantly driven by engineering achievements than by theoretical insights from cognitive science. This certainly does not mean that it is irrelevant to cognitive science" (Millière, 2024, p. 6). Hence, the next section of this chapter will investigate just that – how far the similarities between brains and ANNs allow space for reasonable comparison.

As mentioned, contemporary objections to this comparison are made on the basis that engineering has changed ANNs, distancing them from their origins as brain models. Yet, originally the objection to such comparisons was that the properties of smaller ANNs would not scale up as they grew into multi-layered ANNs (Minsky & Papert, 1969). Then, when PDP networks scaled up, it was instead objected to on the basis that ANNs were sub-symbolic and incapable of modelling language at the level of combinatorial syntactic and semantic structure (Fodor & Pylyshyn, 1988). Which, once again, only lasted until large language models began to demonstrate the ability to model syntax and semantics simultaneously (Buckner, 2019; Elman, 1990; Piantadosi, 2023). Consequently, when we consider the engineering objection more closely, there are several reasons to be doubtful of it. To begin, as the history demonstrates, there is an intuitive resistance to comparisons between brains and ANNs. For, when one such objection does not stand the test of time, another is along shortly to take its place. The content of the resistance may change, but the intuition that meaningful comparisons are unjustified does not. Additionally, the strength of the claim made by such objections has also had to give up significant ground in the face of ANN successes. For instance, early criticisms that ANN capabilities would not scale up and perform the same kind of tasks as brains has wilted and been replaced by the criticism that ANNs are not doing the task in precisely the same way as brains – which is a narrower claim. This persistent retreat indicates

two pieces of information. Firstly, the potential causal powers of connectionist architecture represented by ANNs have not been fully realized until such retreats find a stable position. Secondly, the persistence of the resistance to comparisons between brains and ANNs indicates, whether ultimately justified or not, a thread of brain-exceptionalism as a force behind these objections. This section is not concerned with fully realised connectionist architecture or intuitions about brain-exceptionalism. The following analysis is concerned only with the kind of comparison which contemporary evidence justifies.

2.3.1 Examples from biomimicry

To begin, the brain, like most other products of the natural world, is a product of engineering itself. Engineering in this regard can apply to both people and the natural world as both are responsible for design and creation. Over the course of the last few billion years of life on earth "nature has already evolved and solved many of its problems … Animals, plants and other organisms have engineered themselves to survive and thrive" (Nkandu & Alibaba, 2018, p. 1). The brain is no exception to this process of natural engineering, and it is on this basis the claim that engineering will make ANNs less like brains has good grounds to be challenged[32].

Moreover, it is not as if we have had no success imitating the underlying principles of natural engineering before. Biomimicry is a burgeoning field with numerous successful examples. One of the most famous instances comes from a Swiss engineer by the name of George Mestral, who was taking his dog hunting in 1948 when the animal "emerged from the bushes covered in burrs. After examining the tiny hooks of the burrs, he discovered a hook system used by the plant to spread seeds … inspired by this, De Mestral created Velcro" (Nkandu & Alibaba, 2018, p. 1). But the specific example I am interested in to make my point is in architecture, with the East Gate Centre in Harare, Zimbabwe. The Centre takes inspiration from termite mounds to naturally keep the building cool.

> The self-cooling mounds of African termites inspired the East gate centre design; it stays regulated year round with dramatically less energy consumption ... This mid-rise building designed by architect Mick Pearce in collaboration with Arup engineers, has no conventional air-conditioning or heating. It uses less than 10% of the energy of a conventional building its size through passive cooling and heating techniques … Termites constantly open and close a series of heating and cooling vents in the mounds

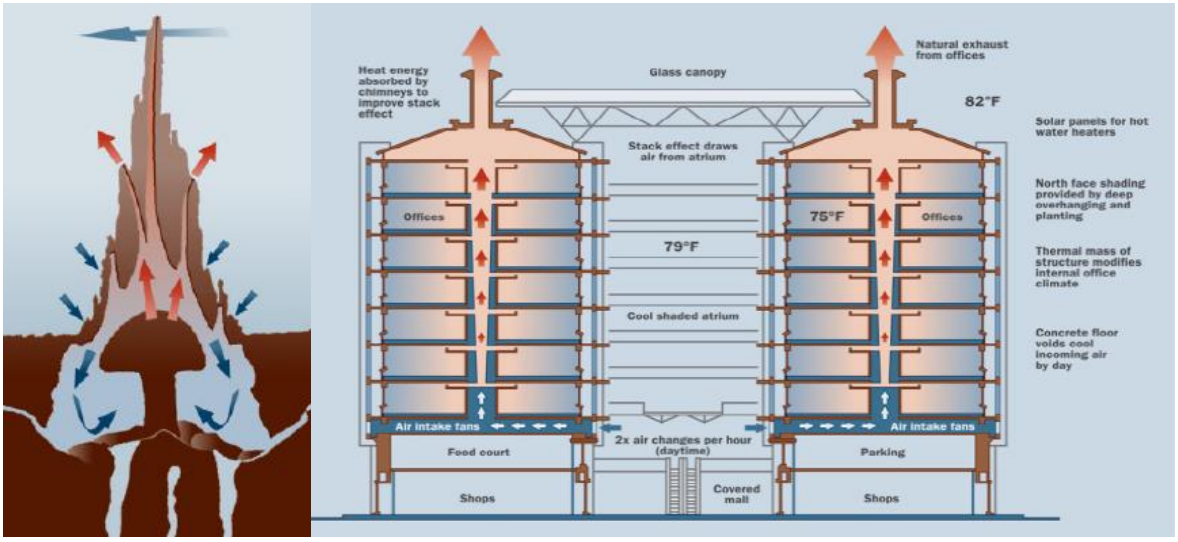
[32] This is not to deny important differences between natural engineering and human engineering. For example, human engineering is top down, while natural engineering is bottom up.

> throughout the course of the day; this keeps the building temperature regulated all day … The East gate building operates in the similar way; outside air is drawn in through vertical ducts on the first floor and is either warmed or cooled by the building mass depending on which is hotter the building concrete or the air. It is then pushed into the building's floors through the central spine of the two buildings before exiting via chimneys at the top (Nkandu & Alibaba, 2018, p. 6).

In this example, the termite mound is an instance of natural engineering within the domain of temperature regulation via ventilation. Its structure and underlying principles have been successfully applied to the East Gate Centre. Though, if one is to compare the East Gate Centre and a termite mound, they are by no means identical. This point is important since the East Gate Centre and a termite mound need not be identical to justify meaningful comparisons. What justifies the comparison is the fact that the underlying ventilation principles have been successfully reverse engineered in the East Gate Centre as a mechanistic abstraction of the ventilation principles of termite mounds. Obvious differences are the kinds of material used, the scale of the respective structures and the composition of the structures themselves. Yet, these differences pale in comparison to the shared underlying principles. In both, air is drawn in from the bottom of the structure, it is then heated or cooled depending on the temperature of the air and the structure, before going through the rest of the structure and finally exiting through chimneys at the top. Below, figure 5, is a diagram of ventilation in the East Gate Centre and in termite mounds.

Figure 7: *(Left) Ventilation in termite mounds (Nkandu & Alibaba, 2018).*

Figure 8: *(Right) Ventilation in the East Gate Centre (Nkandu & Alibaba, 2018).*



Considering this example, the similarities between brains and ANNs can justify a similar comparison. Much like the East Gate Centre or Velcro, ANNs have taken inspiration from something which has already been engineered by nature. As was explored in the earlier parts of this chapter, despite the engineering of the past few decades, the structure, underlying principles, and domains of functional competency remain largely the same in both brains and ANNs. Moreover, these three kinds of similarities are congruent with Marr's three levels - there are similarities in hardware implementation, in the underlying processing principles which transform input to output, and in the kinds of tasks these networks perform well at. The East Gate Centre also shows this, where hardware implementation emulates structural similarities, allowing the instantiation of an underlying principle, which in turn enables the entire building to perform its intended task. Of course, the structure of a termite mound is not necessarily an information processing system[33]. However, it shows the kind of comparison which such similarities enable. Since, considering how the structure of brain networks is preserved in the hardware implementation of ANNs, it is clear that the processing of both is leveraged by the underlying principle of employing high dimensional space to map relationships of statistical dependency spatially. A shared principle which enables brains and ANNs perform well at the same tasks and in turn also justifies a certain kind of comparison. Where the East Gate Centre can be compared as a mechanistic abstraction of the ventilation processes of termite mounds, ANNs can be compared as a mechanistic abstraction of the spatial processing of brains – which opens rich new avenues for analysis.

2.4 Limitations

It can be easy to get carried away with the similarities and make bolder claims than the evidence allows for. As such, it must be reiterated here that brains and ANNs are not the same, and that any comparisons which we do draw between the two must be tempered with an understanding of the limitations of such an analogy. The first powerful objection to these kinds of analogies remains the poverty of stimulus argument (POS). As Chomsky puts it, "our knowledge is richly articulated and shared with others from the same speech community, whereas the data available are too much impoverished to determine it by any general procedure of induction, generalization, analogy, association, or whatever" (Chomsky, 1986, p. 55). Which is to say, the POS argument asserts that children are not exposed to enough language samples[34] to give

[33] Though a termite colony itself may have grounds to be considered this way.

[34] As well as enough samples of the appropriate type. The type of the samples is important too.

a satisfying account of the complex kind of language acquisition humans demonstrate. Consequently, there must be some innate component to language learning such as a UG (Chomsky, 1957; J. Fodor & Pylyshyn, 1988; Pinker & Prince, 1988). Applied to ANNs, the incredible pattern recognition powers demonstrated by these networks are only possible when they are fed enormous amounts of raw data in the training process. This is an important objection because many argue that a human brain can learn using a significantly smaller amount of data. In other words:

> [t]he training regime used in the most successful deep neural models is a further source of implausibility. Models for vision are typically trained with millions of static images, and as many as a thousand images for each category … This amount of data is probably less than the equivalent experience of an infant, but there is an important difference. Once having acquired a basic knowledge of the visual environment, humans are able to learn new categories of objects and actions with a very small number of examples, and can continue to learn all their lives. These natural forms of learning are difficult to achieve with deep neural models … and far distant from the standard training regimes used in the top vision models (Perconti & Plebe, 2020, p. 7).

This discrepancy, unless solved with the creation of far more efficient and generalisable networks in the future will ensure that there remains a gulf between brains and ANNs. Yet, it may also indicate that brains are simply better engineered for the moment, and with further engineering the problem raised by the POS argument may dissolve under the weight of ANN successes. Piantadosi makes this point:

> [i]t will be important to see … how well they [ANNs] can do on human-sized datasets … many of the learnability arguments were supposed to be mathematical and precise, going back to … 1967. It's not that we don't know the right learning mechanism; it's supposed to be that it can be proven none exists. Even my own generative syntax textbook from undergraduate syntax purports to show a "proof" that because infinite, productive systems cannot be learned, parts of syntax must be innate. Proof of the impossibility of learning in an unrestricted space was supposed to be the power of this approach. It turned out to be wrong (Piantadosi, 2023, p. 19).

Although Piantadosi is talking more specifically about Gold's Theorem within computer science in relation to grammar and learnability (Gold, 1967), his point bears on the POS argument too, which has changed over the last few decades. The original criticism was that

learning in this particular way was simply impossible, whereas today the criticism is that it will be impossible from such a small dataset.

While Piantadosi's response to the POS argument is interesting, for the moment it remains speculative that ANNs will solve these problems. Projects such as the *BabyLM Challenge* are an attempt to address this discrepancy, but evidence available today does not rule decisively in either direction. For instance, researchers at New York University (NYU) have attached a camera to a baby to simulate an appropriately impoverished learning environment. Early indications are that: even with a small learning model the input from a single child can result in the learning of things like word-referent pairs (Vong et al., 2024). This research is still very much in progress, and while ANNs may solve the new criticisms being directed towards them by succeeding and training networks of greater sophistication on increasingly impoverished datasets, such is not the case today.

Yet, this is far from the only limitation. It is not out of the question that we are simply unable to train these networks on such small datasets because of some missing feature of composition. Afterall, brains and ANNs are bound by different constraints – both in terms of composition and in terms of the peculiar limitations imposed on biological organisms due to evolution (Hasson et al., 2020, p. 424). In terms of evolution, an ANN can be made far larger than a brain in raw number of units and in raw number of connections. Where evolution may be limited by the forces of natural selection, resources, survival, and environmental pressure, our own construction of ANNs need not be limited by such factors. For instance, ANNs do not have to dedicate processing resources to "learning to balance the body while walking across the room … [or] learning to coordinate hand and finger movements to bring food to the mouth" (Hasson et al., 2020, p. 424). In terms of chemical composition, as previously mentioned brains and ANNs are significantly different in the actual atomic hardware their structures are instantiated on - where one is biological, the other is mechanical, where one is made of carbon, hydrogen, etc. the other is made of metal, silicone, etc. This question of substrate independence may prove to be a problem for any comparison because some of the properties of the brain may be unique not just to the structure, but to the atomic hardware the structure is instantiated on.

For instance, consider photosynthesis. Recent research has found that the process of turning sunlight into chemical energy is not simply happening at the atomic level, but that the process is reliant on events taking place at the quantum level too.

> Photosynthetic organisms can transform the energy contained in sunlight into chemical energy. Pigment molecules - chlorophylls and carotenoids - are excited upon light absorption. The light energy is then transferred, in the form of electronic excitation, in a series of steps to the photosynthetic reaction center (RC), where it is used to induce a charge transfer and, thus, generate an electrostatic potential. This potential is used for further chemical reactions leading to the synthesis of ATP, which provides the free energy for all metabolic reactions in a cell (Ritz et al., 2002, p. 243).

In simple terms, not only does the process of photosynthesis require light sensitive molecules to be able to absorb a photon of light, but it also requires the process of absorption to result in an electron from the molecule becoming excited. Then, the excited electron requires nearby molecules to be capable of transporting it to the organism's photosystem. Once in the photosystem, the electrons need a structure which allows them to spin and release energy as they move, which in turn must generate electrostatic potential, eventually leading to the synthesis of chemical energy in the form of ATP. While this is an oversimplification, the process of photosynthesis demonstrates how reliant complex systems are on an incredible number of particular components arranged in a highly specific fashion. If the light sensitive molecules do not absorb photons, if the absorption does not excite an electron, if the electron is not transported to the photosystem once excited, etc., the system cannot transform light energy into chemical energy. Simply put, if any one of the essential components does not perform its function, the entire system fails. Hence, the point is that such a complex network of contingency may well be unique to a particular biochemical composition, and true reproducibility may be limited to only that which shares its biochemical make up – a line of reasoning which may well be true for the brain too. Searle puts it this way:

> AI has had little to tell us about thinking … Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not (Searle, 1980, p. 424).

The limitation here is that the properties themselves may not be independent of chemical composition. In the case of photosynthesis, perhaps this is the only composition of atoms which will result in light energy being successfully transformed and stored as chemical energy. If this is true, no other system will suffice. Applied to brains, perhaps instantiating the full causal powers of the brain on a system of silicon, metal, etc. will simply not suffice too. In this regard, the limitation may be that ANNs are able to generate a mechanistic abstraction of the spatial processing of brains, but much more than that will be impossible because properties like consciousness, intentionality, etc. are reliant on the especial biochemical composition of the brain, i.e. these properties may not be substrate independent.

This view even has some support within the current literature. For example, a recent study in the quantum behaviour of microtubules has found evidence to support a decades old hypothesis about consciousness. Originally put forward by physicist and mathematician Roger Penrose, he was not convinced that consciousness could be achieved by a traditional, algorithmic computation and theorised that its non-reducibility was the result of the brains reliance on some kind of quantum phenomenon (Penrose, 1989). Initially he was unsure about exactly where the brain was engaging in quantum processes, but Stuart Hameroff would point him in the direction of microtubules (Penrose, 1994). This collaboration culminated in the Penrose-Hameroff model of consciousness (Hameroff, 1998), which would be criticised on the basis of quantum decoherence and the prevailing suspicion that the brain was simply too slow to do any kind of quantum computation (Tegmark, 2000). Evidence from 2024 has contradicted that view by finding quantum effects within the microtubules themselves. The study observed:

> Trp [Tryptophan] UV-excited transition dipoles in microtubule architectures, which leads to an enhancement of the fluorescence quantum yield (QY) that is confirmed by our experiments…. which is believed to be the first UV light perception system discovered to use a network of Trp chromophores as a funnel to enhance its quantum efficiency … Therefore, our work demonstrates that collective and cooperative UV excitations in Trp mega-networks support robust quantum states in protein aggregates (Babcock et al., 2024, pp. 4035–4044).

However, the findings of this experiment remain far from saying anything about consciousness just yet. The focus was limited to quantum effects regarding light, and the boldest claim made regarding brains and quantum physics was that the microtubules may serve “a photoprotective role in pathological conditions such as Alzheimer’s disease and related dementias” (Babcock

et al., 2024, p. 4044). Yet, these findings point us in the direction of Searle's argument once again. Considering the possibility that the brain relies on quantum effects generated by its biological composition and architecture, its unique properties may well "be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena" (Searle, 1980, p. 424). Nevertheless, this limitation remains hypothetical, and does not itself provide reasons to think that these quantum phenomena could not be reproduced in another system comprised of metal, silicon, etc. Consequently, our analogy must remain tempered by the reality of what we see in front of us right now.

What we see in front of us right now is that no ANN is identical to a brain, nor does any ANN replicate the full causal powers of one or indicate the imminent arrival of AGI. Instead, ANNs demonstrate remarkable similarity when we compare them to brains as information processing systems. When compared as information processing systems, we find that the basic structure of brains is preserved in ANNs. This is a similarity at the level of hardware implementation. Next, at the level of representation and algorithm, we find that the basic underlying principles of processing in brains are also largely preserved in ANNs, as the processing of both is leveraged by using high dimensional space to represent relationships of statistical dependency spatially, representations which take the form of activation patterns. Finally, brains and ANNs perform well at similar tasks. These common domain competencies include, but are not limited to visual processing, auditory processing, language processing, as well as generation of art, music, etc. This is a similarity at the level of computational theory insofar as they are aimed at the same goals and largely use the same strategy to achieve them. Therefore, while ANNs are by no means identical to brains, they have enough in common to be compared as a mechanistic abstraction of the spatial processing of brains – a comparison which opens rich new avenues for understanding ANNs.

3. Something Emergent This Way Comes

Despite the similarities from the previous chapter, some may still object to the comparison that ANNs serve as even a mechanistic abstraction of the spatial processing of brains, insisting "on the irrelevance of deep learning models for cognition, imputing the analogies between patterns in the brain and models to mere coincidence. But insisting on this line of reasoning is vulnerable to a sort of 'no–miracle argument'" (Perconti & Plebe, 2020, p. 10). The no-miracle-argument, originally applied by Putnam to scientific realism, points out that our success in observing and recreating natural phenomena, like in the natural sciences, is more than enough evidence to justify a realist position towards their findings. This is because we can hold reasonable doubt about whether these sciences are mapping reality in a way which is complete, while still taking the implications of their findings seriously. In the context of his argument, Putnam explains:

> a natural explanation of the success of these theories is that they are partially true accounts … And a natural account of the way in which scientific theories succeed each other - say, the way in which Einstein's Relativity succeeded Newton's Universal Gravitation - is that a partially correct/partially incorrect account of a theoretical object … is replaced by a better account of the same object or objects. But if these objects don't really exist at all, then it is a miracle that a theory which speaks of gravitational action at a distance successfully predicts phenomena … [Hence] my interest is rather in the following fact: the realist's argument turns on the success of science … [and] [t]hat science succeeds in making many true predictions, devising better ways of controlling nature, etc., is an undoubted empirical fact (Putnam, 1978, p. 19).

So, even if the findings of the natural sciences are not mapping reality exhaustively, what they do succeed in, is, through their successes, identifying real objects of interest and relationships between real phenomena. To object to this, says Putnam, is to argue for the miracle, that the persistent and precise findings of science are, in fact, better explained as mere coincidence.

Those who object to even the weaker claim for comparison (that ANNs are a mechanistic abstraction of the spatial processing of brains) find themselves in a similar position. Since, if they are to follow this line of reasoning to its conclusion, they too must argue for a miracle. Accepting, that in the face of the variety and significance of similarities present across all of Marr's levels, the best explanation for the persistence and precision of the performance similarities between brains and ANNs is, again, by some miracle, better explained as coincidence. Of course, since the first chapter established that ANNs were designed as an

attempt to mimic how brains process information, we know this is not the case. Accordingly, the aim of the final chapter is to investigate the consequences of accepting the weaker claim. In particular, the ramifications of this claim for epistemic opacity in ANNs. Since, if they are a mechanistic abstraction of the spatial processing of brains, an explanation more aligned with current evidence is that epistemic opacity is an emergent feature of ANNs, not an epistemological bug[35].

In terms of emergence in ANNs, the explanation is going to consist of a threefold scaffolding. First, and most importantly, ANNs fulfill the criteria for emergence. Second, building on the foundations laid by the first two chapters, I argue the history of ANNs has long indicated that emergent properties would eventually arise. ANNs did, after all, have their origins in modelling the brain. Hence, taking inspiration from the structure and processing principles of brains serves as a teleological force for emergent properties to arise. Since, if the brain possesses emergent properties and we attempt to imitate its causal powers through the creation of a mechanistic abstraction of its spatial processing, successful attempts will tend to generate the same epistemically challenging phenomena we encounter with brains: emergent properties. Third, work by Hasson et al. (2020) supports this position by highlighting that brains and ANNs are engaged in the same kind of processing, interpolation; and that ANN engineering has even converged on some of the same processing solutions that brain evolution discovered. Together, these arguments demonstrate not only that ANNs fulfill the criteria for emergence, but that there are teleological and structural forces which brought this about and make its presence unsurprising. Hence, epistemic opacity can be reframed as an emergent feature, not an epistemological bug.

3.1 What is emergence?

While the philosophical notion of emergence is entangled with various debates concerning matters such as downward causation (Kim, 1999), the degree to which emergent phenomena can be reduced to material causes (Kim, 1989, 1999), or attempts to distinguish different types of emergence (Chalmers, 2006), this paper is focused on a broader notion of the concept and will bracket ancillary discussions for the time being. In that regard,

[35] The choice of the word bug is intentional here and I employ it in the sense of computer bug, not biological bug. Computer bugs are generally understood to be design mistakes which lead to incorrect or surprising outcomes. Epistemic opacity in ANNs has come to be seen as something of a computer bug in recent years. I argue against this characterization because its presence is not the result of mistaken design.

> [a]t the core of these ideas [is] the thought that as systems acquire increasingly higher degrees of organizational complexity they begin to exhibit novel properties that in some sense transcend the properties of their constituent parts, and behave in ways that cannot be predicted on the basis of the laws governing simpler systems (Kim, 1999, p. 3).

The above is especially useful to focus our analysis on the broader concept here. What I am referring to is those instances in which the whole takes on properties which its parts do not possess. Hence, the focus of this section is to build a more general criteria to identify emergence. Broadly speaking, "the concept of emergence [is understood] as the unavailability of a certain scientific explanation for an observer or observers" (Taylor, 2015, p. 667). However, emergence is not simply things which we have difficulty generating satisfying scientific explanations for, emergence also requires that the explanatory issue be of a certain kind. According to Wimsatt, that which we specifically have trouble explaining about emergence stems from a failure of aggregativity (Wimsatt, 1997, p. 382). In other words, when it comes to a system with emergent properties, the features we tend to attribute to these systems are twofold. The sum of the systems parts does not add up to the whole, and aggregative accounts cannot adequately explain why. This conception is consistent with the broader view of emergence held by Kim, Wimsatt, and even Chalmers. Though, Kim and Chalmers may object to some of the details here.

Accordingly, there are two general criteria for assessing emergent phenomena. Firstly, scientific explanations based in observation fall short of resolving the emergent phenomena. Secondly, there is a failure of aggregativity between the systems constituents and higher-level system properties – i.e. when we add up the constituent parts (the micro-level) they do not aggregate to the whole (the macro-level). This means that failures of aggregativity are characterised by the macro-level possessing properties which are not present at the micro-level. Though, as Wimsatt points out, this definition may not be adequate for some, since "very few system properties are aggregative, suggesting that emergence, defined as failure of aggregativity, is extremely common - the rule, rather than the exception" (Wimsatt, 1997, p. 382). This objection is not necessarily a problem though. Simply saying something is common does not necessitate that it is not emergent. Which is part of Wimsatt's point – that, perhaps the regularity with which aggregativity fails and at which emergent properties become detectable, indicates that emergence, at least understood as a failure of aggregativity, is far more common than we realize. Reframed this way, failures of aggregativity are not the problem they are often made out to be. Instead, they are powerful markers which give us an indication,

not only of where the gaps in our understanding reside, but of where to look if we are to fill them in. Wimsatt himself puts it like this, "[w]e are too much taken with things which fit our models, but often have more to learn from edges that do not fit … [In this way] [i]nvariance claims are particularly sharp-edged tools" (Wimsatt, 1997, p. 383). More specifically, failures of aggregativity are tools in-so-far as they serve as markers which indicate the presence of emergence. Once such failures are identified we are a step closer to understanding the emergent phenomenon itself. Since, we now at least know where to look. Though, the inclusion of our first criteria, the lack of satisfying scientific explanations, is a component which remains sensitive to critics of Wimsatt who argue his conception is too broad.

Of course, Chalmers may challenge such a broad notion on different grounds, arguing that it does not distinguish types of emergence. His distinction between strong emergence (henceforth strong[e]) and weak emergence (henceforth weak[e]) may be raised as an objection. An objection of this kind would likely argue that there is significant enough difference between the two types of emergence to prevent a consolidated criteria of the kind I am attempting to generate. Yet, it is possible to acknowledge the distinctness of these two types without decoupling them entirely. Indeed, by the definitions he provides, both strong[e] and weak[e] appear to be aimed at a larger shared referent. For example,

> [w]e can say that a high-level phenomenon is *strongly emergent* with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are not *deducible* even in principle from truths in the low-level domain … [Whereas] a high-level phenomenon is *weakly* emergent with respect to a low-level domain when the high-level phenomenon arises from the low level domain, but truths concerning that phenomenon are *unexpected* given the principles governing the low level domain (Chalmers, 2006, p. 244).

Here, both strong[e] and weak[e] are still characterised by a difficulty aggregating the parts and the whole. Cases of strong[e] just cannot be explained in principle by looking at the constituents, while weak[e] can. Hence, with weak[e], it is simply a surprise that these constituents led to this phenomenon, not an explanatory problem. Importantly, while both would constitute a failure of aggregativity insofar as the sum of the parts does not aggregatively add up to the whole, only one involves an epistemically serious problem for explanation. Here my criteria from earlier in section 3.1 become relevant again. Since the criteria enable the definition to remain broad while still being sensitive to important debates in the field of

emergence. For example, strong[e] would involve the fulfilment of both criteria: (1) scientific explanations do not adequately explain the phenomena and (2) there is a failure of aggregativity between the parts and the whole. On the other hand, weak[e] would fulfil (2), but not (1).

In being sensitive to this distinction, we encounter the kind of objection Kim can level against my broader criteria. When Chalmers pointed out that some kinds of emergence may not, even in principle, be reductively explained to generate a satisfying account of how the material constituents come to produce higher level properties, he touches on Kim's core characterisation of the concept itself. Since Kim is of the position that "emergentism is a form of what is now standardly called 'nonreductive materialism', a doctrine that aspires to position itself as a compromise between physicalist reductionism and all-out dualisms" (Kim, 1999, p. 4). This means that emergent properties may have material causes, but not reduce to material explanations. Kim has argued in more detail elsewhere against such non-reductive materialism (Kim, 1989); his argument is that such a "halfway house is an inherently unstable position, and that it threatens to collapse into either reductionism or more serious forms of dualism" (Kim, 1999, p. 5). Again, for the purposes of this paper we will bracket ancillary discussions on whether satisfying accounts of emergence can be exhausted by non-reductive materialism alone[36]. I merely bring it up to highlight that my criteria can account for Kim's position. Since, he would consider weak[e] to not be emergence and strong[e] to be emergence. Hence, even if he was unhappy with the broadness, his position still exists in this account as the fulfilment of both criteria (1) and (2) simultaneously.

Considering all the above, the criteria I have established are well placed to serve as a middle ground between some of the more extreme positions in the field. Such a framing does not need to rule on matters regarding downward causation, degree of material reducibility, which kind of emergence is 'truly' emergent, etc. It merely seeks to provide a description of the two elements we tend to take most seriously in any good theory of emergence: (1) difficulty providing satisfying scientific explanations, and (2) failures of aggregativity between constituents and high-level system properties[37]. Therefore, regardless of one's position, these criteria can capture both our weakest and strongest conceptions of emergence.

[36] Especially since accounts objecting to reductive materialism have similarly serious objections of their own, such as those present in Fodor's *Special Sciences* (1974).

[37] Additionally, consider each of these criteria to be scalar depending on the relevant position.

Of course, the weaker version will likely not satisfy serious critics, so we will have to show, contrary to the view of Chalmers (2006, p. 252), that connectionist networks, especially those with newer architectures, actually do fulfil the criteria for both (1) and (2), making them an example of strong[e], not weak[e]. In terms of criteria (1) the epistemic opacity present in ANNs is a classic example of difficulty providing satisfying scientific explanations. In terms of criteria (2), more detailed analysis of the opacity itself will point towards precisely where these difficulties in explanation are coming from – a failure of aggregativity arising between constituent parts and higher-level properties.

3.2 What is epistemic opacity?

Epistemic opacity in ANNs, known popularly as the 'black box problem', is marked by our difficulty generating satisfying explanations for why these networks come to arrive at their 'decisions'. Simply put:

> in their current incarnation, deep learning systems have millions or even billions of parameters, identifiable to their developers not in terms of the sort of human interpretable labels that canonical programmers use ("last_character_typed") but only in terms of their geography within a complex network … [as such] neural networks as a whole remain something of a black box (Marcus, 2018, pp. 10, 11).

Though, such references to the sheer number of parameters involved are only a component of why we have difficulty understanding ANNs. Take, for instance, the larger machines from the classical computationalist school of AI - machines like Deep Blue II (the first computer to beat a world chess champion). It was incredibly large and complex insofar as it could search up to 330 million chess positions in a second, and it had a sophisticated search tree structure (Campbell et al., 2002). Yet, when compared to ANNs, the decisions of this network remain straightforward and interpretable despite the sheer amount of information being processed. It is built on the principles of classical computationalism, and it works on a system of if-then rules. Since it relies on these rules, this means that it also produces code that tells the programmers what decisions it took, and which rules determined these decisions. An objection could certainly be made that the sheer amount of code one would have to filter through to understand all the decisions still makes interpretability difficult (after all, Deep Blue II's processing involves trillions of searches for correct chess moves), but there is still a direct and straightforward way we can understand the computers move selection. In other words, the code

produced by these machines, is massive in scale, but remains intelligible and there is recourse to understanding its internal reasoning. This is not the case with ANNs:

> there is a straightforwardly mathematical sense in which deep neural networks are fully transparent. All weights on all connections across the network, billions as there may be, are both available to inspection and computationally tractable. However, while formally precise, neural network logic is largely semantically unintelligible. That is, the mathematical expression of a fully trained neural network model cannot, in general, be given an intelligible interpretation in terms of the target system such that one can understand or comprehend how the parts interact and contribute to the networks' outputs (Duede, 2022, p. 1091).

Thus, the problem with ANNs is not just one of scale, but that the nature of the processing itself makes intelligibility nearly impossible. Factors like the highly distributed manner in which the parts interact, or the influence of gradient descent techniques on the intelligibility of the whole system give rise to outputs which remain elusive despite our best efforts. Perhaps in time the whole may be calculable from the parts, but for now, the very processing of these networks appears to be characterised by a failure of aggregativity. In this way, the breadth of the opacity indicates that its presence is not a bug in the system but a feature of such networks - an opacity by natural design if you will[38]. And the significance of the difficulty explaining the internal reasoning of ANNs benefits from placement in real world contexts. For example,

> [i]f an AI system tells us that Lucie should not get a mortgage, she is entitled to understand why she should not get a mortgage. To answer the why question by simply insisting that the decision was made by a reliable but incomprehensible algorithm isn't good enough ... She is entitled to receive a justifying reason for the rejection … [since the answer has] revealed nothing of the internal reasoning that might have gone into producing that output (Cappelen & Dever, 2022, pp. 162 & 163).

This example hones in on the problem Duede identified on the previous page. If we cannot understand how the constituent parts interact and contribute to the output of *Lucy should not get a mortgage*, then we do not understand the internal reasoning of the ANN and consequently cannot even give Lucy a reason for the output. Reasons like the one mentioned by Cappelen and Dever, such as 'the algorithm is reliable', are reasons to accept the output, but not reasons

[38] The details of why the opacity is by 'natural design' will be picked up in the next section.

for how the output was reached in the first place. The difficulty in generating the second kind of reason originates from the kind of information processing systems ANNs are. They transform information into a function by distributing the processing load throughout the network, and in doing so, break semantically intelligible information into fragments[39]. As covered in the previous chapter this distribution of the information allows the network to map the inputs within its own high dimensional space and represent relationships of statistical dependency with use of activation patterns (Elman, 1990), but this also generates opacity in a specific way – through what I call fragmentation, amalgamation and reduction processing (FAR processing henceforth).

The three major components of FAR processing are as follows. First, fragmentation - discrete pieces of information are broken up into fragments and distributed throughout the network. Second, amalgamation – the information is not just broken up into fragments, these fragments are also combined with other fragments from different pieces of once discrete information, making intelligibility even more challenging. Finally, this process of fragmentation and amalgamation culminates in a reductive output being transmitted to the next layer (Krenker et al., 2011, p. 3). In other words, the variety of discrete fractions of information amalgamated in the artificial neuron are then reduced to an output consisting of either a binary or scalar representation of its activation. Regardless of whether this output is binary or scalar, it representants an information bottleneck because the full complexity of the information is not preserved by either scalar or binary activation of the local unit. In this way, all three factors contribute to the overall result - satisfying explanations of ANNs internal reasoning is rendered epistemically challenging. Since, by the time the network transforms the information back into something semantically recognizable, like those larger processing units we call activation patterns, the internal reasoning which led to those patterns has already been lost by distributing the load across the high dimensional space of the network and processing information in this reductionist fashion. Crucially, the previously discrete units of information have already been unintelligibly fragmented, amalgamated and reduced before becoming intelligible again. Therefore, even if activation patterns may be represented as functions, and even if these functions take on semantic significance, the internal reasoning which led to the generation of these functions remains opaque. There are methods like compression which aim to simplify ANNs while minimizing performance sacrifices (Yao et al., 2021), as well brain inspired ablation techniques (Meyes et al., 2019). While these attempts can lead to answers, they can

[39] Whether these fragments are intelligible or not is still a contested part of the literature (Calegari et al., 2020).

also lead to more questions. For example, some features tend to be continuously represented in the same parts of the network despite ablation, while for others this is not the case and structural damage to the network results in some features being represented in different local units (Meyes et al., 2019, p. 15). This kind of flexibility makes these networks incredibly robust, but it also poses a serious problem for explainability.

Considering these kinds of problems within the context of emergence, epistemic opacity in ANNs can now be reframed. For, ANNs fulfil the two basic criteria for emergence. Firstly, the presence of epistemic opacity discussed here is a clear indication that scientific explanations based in observation fall short of giving a satisfying account of the higher-level properties which these networks exhibit. Secondly, what these explanations are specifically unable to explain arises from a failure of aggregativity. The bare fact that "a neural network model cannot, in general, be given an intelligible interpretation in terms of the target system such that one can understand or comprehend how the parts interact and contribute to the networks' outputs" (Duede, 2022, p. 1091), is by definition a failure of aggregativity. Since, when we aggregate the constituent parts, what we fail to generate, is an adequate account of how the system came to possess the properties which it appears to have[40]. For an illustration of this, think back to the case of Lucy given to us by Cappelen and Dever. We are unable to explain why the ANN produced the output that she should not receive a mortgage, and this failure of explanation is the result of the same kind of explanatory gap. The kinds of reasons which we attempt to give are that the ANN is reliable enough for us accept the resulting output. This may be a good reason for a banker, but it is not a good reason for an epistemologist. There remains an explanatory gap between the constituent parts and how the relationships between them result in the output of the network. Aggregativity fails to bridge this gap. More precisely, the gap between constituents like artificial neurons, weighted connections, etc. and higher-level properties such as: the pattern recognition powers which identify Lucy as a high-risk mortgage candidate, clearly represent a failure in aggregativity because, when we simply sum these parts, such aggregation does not explain the presence of the higher-level properties. This kind of invariance is what Wimsatt proposed as a tool to indicate the presence of emergence. Here is an edge which does not fit our model. Here is an instance where summing the parts does not result in a satisfying explanation. Here is a failure of aggregativity. Consequently, at least according to Wimsatt's conception, ANNs fulfil our basic criteria for emergence.

[40] Though, the failure of aggregation could have a variety of causes, such as parameter complexity.

That ANNs fulfil these criteria is only a component of my argument. The forthcoming parts of this chapter advance two more reasons in favour of understanding epistemic opacity in ANNs as an emergent feature. The first, is that the historical origins of ANNs set up a teleological force which made any successful creation likely to occur with emergent properties. The second is that there is a growing body of contemporary evidence which supports my argument. For, not only do brains and ANNs engage in the same kind of processing, interpolation, but they have converged on some of the same interpolative processing solutions as brain evolution too.

3.3 ANNs are emergent because brains are emergent

The first intimation that ANNs could come to possess emergent properties was the Perceptron (Rosenblatt, 1958). Its basic structure led to the emergence of more complex properties and attracted substantial media attention, but for the wrong reasons. The popular sentiment in many articles of the time was that 'the network recognized shapes in the way humans do'. While such specifics may have been hyperbolic, the underlying sentiment behind the enthusiasm did have credibility. Something new had come from simple constituents. Pattern recognition powers which were not present in any of its parts had emerged at the level of the system. The Perceptron could 'recognize' shapes and, as I will argue, this came with the beginning of emergent properties. Moreover, since Rosenblatt was candid about where his inspiration had come from, the success of these early ANNs can be directly connected to their relationship with brain science. Rosenblatt was, after all, a psychologist attempting to replicate the causal powers of the human brain. Which is to say that the Perceptron was a rudimentary kind of brain model directed at the problem of photosensitivity. Yet, with the arrival of heavy criticism and the engineering of ANNs, such comparisons were lost for a time (Boden, 2016). The successes of the newer networks are making the parallels hard to ignore much longer, and the presence of epistemic opacity in these networks provides a powerful link between brains and ANNs. For, it seems that the processing of brains is similarly opaque to the processing of ANNs. More specifically, the distributed spatial manner in which both networks process information seems to be what generates this epistemic difficulty. Significantly though, we refer to our difficulty explaining such failures of aggregativity in brains as the problem of emergence, while in ANNs we refer to it as epistemic opacity. For instance, it is uncontroversial that the brain is rich in a variety of emergent features. Among many others,

> [l]anguage, thought and mind are the ultimate cases of emergence. Consciousness, the first-person perspective, our sense of time passing and our awareness of our own

> consciousness are the most elusive objects of scientific understanding. However, the brain also produces an extraordinary range of highly sophisticated emergent features, most of which are sub- or unconscious. For example, reaching out to turn a door handle, face recognition and turning the head towards a sound are all very high-level processes that result from the firing of specialised neurons in specialised regions of the brain (Ladyman & Wiesner, 2020, p. 60).

Here, it is important to limit our focus to a specific set of examples of emergence in brains. As mentioned previously, my claim is not that these networks model the mind, but that they are a mechanistic abstraction of a particular component of neural processing - the distributed, spatial kind. Hence, my claim is only relevant to some cases of emergence and has little to say about the higher-level emergent properties associated with minds - thought, consciousness, intentionality, etc. There would need to be examples of ANNs creating these kinds of emergent properties before we could reasonably consider them as a candidate for causal responsibility. Therefore, the weaker claim is only relevant to examples of emergence present in both brains and ANNs already - like art and language generation as well as audio processing and visual processing.

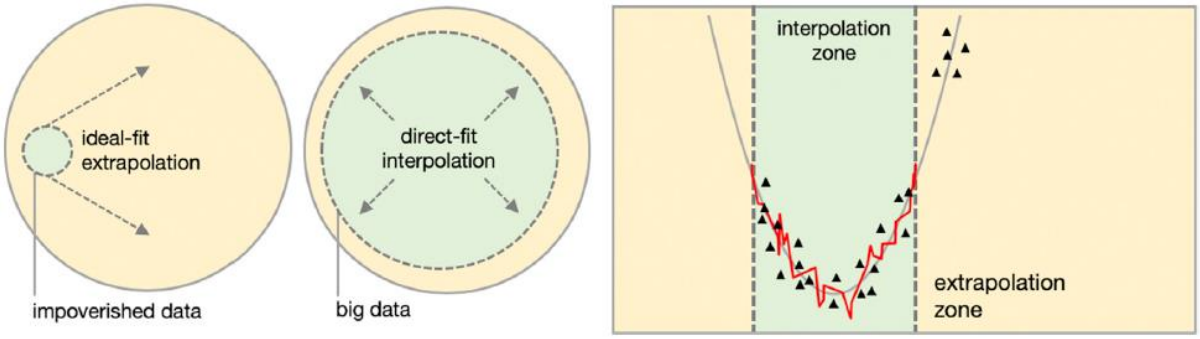
3.3.1 Convergent processing solutions

Reframing these epistemic difficulties as an emergent feature shared with brains has a growing body of evidence to support it. For, the teleological force I mentioned earlier in the chapter is grounded in the idea that many of the problems related to distributed spatial processing have pre-existing, well optimized solutions in the form of brain evolution. In this way, even when ANNs left the fields of philosophy, psychology, cognitive science, etc. for engineering, physics, mathematics, etc. it may not have altered their trajectory enough to prevent meaningful comparisons. Since, with ANNs being mechanistic abstractions of the distributed, spatial processing of brains, they may well encounter many of the same processing problems and processing solutions. Hence, founding ANNs as a model of a brain may have doomed them to solve the same problems. New research by Hasson et al. “investigates how ANNs learn to perform complex cognitive tasks and whether the solution is at all relevant to cognitive neuroscientists” (2020, p. 416). They conclude that it is relevant, and that the evidence does in fact support convergence on the same solutions for brains and ANNs. Their evidence is that both brains and ANNs are direct-fit models that rely on the same kind of algorithmic interpolative techniques. For,

> [a]cross species and models, BNNs [biological neural networks] and ANNs can differ considerably in their circuit architecture, learning rules, and objective functions … All networks, however, use an iterative optimization process to pursue an objective, given their input or environment - a process we refer to as direct fit (Hasson et al., 2020, p. 416).

This kind of direct-fit modelling is contrasted with ideal-fit models. Ideal-fit models achieve their ends through learning "the underlying generative structure of the data by exposing a few latent factors or rules" (Hasson et al., 2020, p. 418), meanwhile direct-fit models do not. Instead, direct-fit models pursue their objective by directly mapping what is in their environment (for brains) or based on the data input (for ANNs). Until recently, interpolative, direct-fitting methods were widely viewed as weaker at generalization and prediction tasks; leading us to believe that the brain was performing tasks through extrapolation[41] rather than interpolation[42]. Yet, big data and big networks have turned this assumption on its head (Hasson et al., 2020, p. 419). For example, modern ANNs have become so good at predictive tasks, that it is easy to misinterpret them as an ideal-fit model doing extrapolation. But, to the contrary, newer ANNs have become remarkably effective at such prediction tasks by doing interpolative direct-fitting; what changed, at least according to Hasson et al., was the size of their interpolation space. Increasing the area of interpolation improves their predictive powers because "within the interpolation zone, the ANN is as good as the ideal-fit model in predicting the values of new observations not seen during training" (Hasson et al., 2020, p. 420).

Figure 9: *Differences between extrapolation & interpolation (Hasson et al., 2020, p. 419).*



[41] Extrapolation is a data analysis technique which attempts to predict new values outside of the range of its data set by assuming that trends in the current set will continue to cases outside of it.

[42] Interpolation is a data analysis technique which attempts to predict new values within the range of its data set by assuming that the referential relations around a new value will allow accurate estimations of it.

Therefore, as you increase the interpolation zone, you also increase the range of the network's predictive powers. However, since "[t]he density and diversity of training examples determine[s] the interpolation zone" (Hasson et al., 2020, p. 426), to increase this area, the networks need more information and more parameters to fit the millions or even billions of examples which constitute the interpolation zone. This means that, generally[43], bigger networks with access to more data have a bigger interpolation zone and greater predictive powers.

Looking back at chapter 1, the smaller interpolation zone was a problem during the first two epochs of ANN development, but not the third. With the engineering paradigm, the size and sophistication of ANNs improved rapidly along with better GPU technology and the arrival of big data. Meaning that these networks now had the high dimensional space to generate an interpolation zone with enough parameters to be useful for prediction tasks. At the same time, big data provided the other component necessary for the growth of the interpolative space in direct-fit models, the millions of examples needed to take advantage of their increased parameterization. Hasson et al. make the claim that these two conditions for improved interpolation are solutions which the brain, as a direct-fitting model, has already converged upon. First, they point out that, like modern ANNs, brain processing is computationally rich and overparameterized. Their evidence is that processing resources are not exactly a scarcity in the brain, arguing that:

> [e]ach cubic millimeter of cortex contains hundreds of thousands of neurons with millions of adjustable synaptic weights, and BNNs utilize complex circuit motifs hierarchically organized across many poorly understood cortical areas … [Hence, while] the brain is certainly subject to wiring and metabolic constraints, we should not commit to an argument for scarcity (Hasson et al., 2020, p. 428).

Secondly, they are not convinced by the central thesis of the POS argument and believe that the brain's learning environment may be far richer than previously thought. Pointing out that for humans

> [r]ecent measurements suggest that the incoming input may be vast and rich … For example, we may be exposed to thousands of visual exemplars of many daily categories a year, and each category may be sampled at thousands of views in each encounter,

[43] There are specific cases in which the dataset does not need all the parameters it has access to and making use of all of them just increases the time it takes for the computation to run (Buckner, 2019, p. 7).

> resulting in a rich training set for the visual system. Similarly, with regard to language, studies estimate that a child is exposed to several million words per year (Hasson et al., 2020, p. 428).

Considering their evidence, as well as my own in Chapter 2 (Azevedo et al., 2009; Ladyman & Wiesner, 2020; Yan & Hricko, 2017), the first point that the brain is overparameterized and rich in processing resources is quite well supported. However, I am not convinced that their second point, the dismissal of the POS argument, is as well grounded in evidence; especially since current research is still focused on resolving this question (Vong et al., 2024). With that being said, this is not their only evidence against the POS argument, and they also emphasize that for humans some objective functions may be innately stored and that "social exchange provides a basis for supervised learning" (Hasson et al., 2020, p. 424), which would form the context needed to reduce the required information for effective interpolation (Millière, 2024, p. 5). Yet, while we may not have evidence to rule decisively on the POS argument, the richness of resources and parameters available to the brain are strong indicators that its tasks require a richness of information as well. For, as was pointed out with the conditions for the success of ANNs, to have a big enough interpolative zone for strong prediction and generalization, big networks also need massive amounts of information. Hence, the interpolative processing of brains may well require a rich set of examples to draw from. In support of this, the work of Luc et al. points out other reasons why the input brains receive may not be as impoverished as we previously believed. For, when trained on frames from a video sequence instead of static images, a single second of video ends up consisting of 30 different frames (Luc et al., 2017, p. 655). Meaning that, based on those calculations, a day of video input could constitute more than 2 million visual exemplars. Moreover, these calculations do not factor in other forms of sensory input such as sound, touch, taste, smell, etc. or episodic replay during sleep, which may enable a single exemplar to be reused in training numerous times (Buckner, 2019, p. 13). Hence, while the evidence may not rule out the POS argument entirely, for the brain to have a big enough interpolative zone to generalize and predict in the way that it does, it seems it would also need to be much richer in information than the POS argument suggests.

This is precisely the teleological argument I mentioned earlier. Having founded ANNs as a model of the brain which serves as an abstraction of their distributed spatial processing principles, we may have destined them to solve the same problems which brain evolution has

spent eons on[44]. The work of Hasson et al. supports this by pointing out, not only that "ANNs and BNNs belong to the same family of direct-fit models" (Hasson et al., 2020, p. 417), but that this family of direct-fit models are algorithmically reliant on interpolation and have converged on the same solutions to solve problems with prediction and generalization; they both solved these problems by increasing their interpolative zone through the mechanisms of increased parameterisation (bigger networks) and denser, more varied sampling (bigger data). Considering this, it should not come as a surprise that, like brains, understanding ANNs is an epistemically challenging task. For, if they are encountering the same processing problems, and reaching the same solutions to those processing problems, it stands to reason that understanding them will be similarly difficult too. In this way, epistemic opacity in ANNs may not be the kind of system bug which it is generally portrayed as. Instead, it may well be an emergent feature of this kind of complex system. Hence, the cause of the epistemic difficulties has its origins in the same referent too, emergence. Though, on the matter of emergence and epistemic opacity, Hasson et al. do not share my position. For example:

> [they] argue that there is nothing opaque about ANNs - they are fully transparent 'glass boxes' … we deem ANNs [as] black-box models … because we are deeply committed to the assumption that the ANN must learn a set of human-interpretable rules necessary for processing information. This is our classical criterion for understanding. Since we do not readily find such rules when interrogating the distribution of millions of adjustable weights within over-parameterized artificial (and biological) neural networks, we demote such models to black-box status (Hasson et al., 2020, p. 423).

In other words, they are of the position that ANNs are not the kind of things which can be understood in terms of human-interpretable rules or labels, therefore our inability to explain them in such terms is not a problem. Considering the work they have done to point out the interpolative processing, I can see how they acquired the impression that there is in fact no black-box problem. From their point of view these models are not finding any rules which can be extrapolated; the increased interpolative space just allows for more accurate predictions. Hence, the reason they may provide is something like the algorithm has predicted a new example accurately because it was within the interpolation zone. Much like with the Cappelen and Dever's example of Lucy, this is reason to accept the output as reliable, but not an

[44] Of course, a key difference between brains and ANNs is that we explicitly designed them to solve these kinds of problems.

explanation for why that output was reached in the first place. Such interpolative processing may be a brute force over-parameterisation which is reliant on the structure of the world contained in its input, but this does not entail that it cannot be explained further. Hence, their position on the black-box problem appears to be inconsistent with their characterisation of interpolative processing. For, they admit to the existence of a structure to the world which interpolative processing picks out, but insist that it can only be understood at the level of weights, neurons, networks, etc. Yet, if successful interpolation is as reliant on the structure of the world as they argue, then this means that it is picking out real world phenomena that the network is identifying. None of which seems to prevent us from identifying these phenomena in human terms. As Duede puts it:

> despite their epistemic opacity, deep learning models can be used quite effectively in science, not just for pragmatic ends but for genuine discovery and deeper theoretical understanding, as well. This can be accomplished when DLMs are used as guides for exploring promising avenues of pursuit in the context of discovery (Duede, 2022, p. 1097).

In this way, there is a gap between how well we have optimized these networks to pick out interesting phenomena in the world through pattern recognition, and how well we understand the internal reasoning which allows them to pick these phenomena out in the first place. Hence, the networks are only 'discovering' phenomena because of the structured relations that exist in the world - which is why Duede suggests that we use them as markers of where to direct further scientific efforts. Yet, if what these networks are helping us to discover can be explained as Duede expects, then the internal reasoning which picked them out will also have some explanation which goes beyond weights, neurons, networks, etc. Consequently, contrary to Hasson et al., the black problem seems like it will remain until we are able to explain the internal reasoning of ANNs beyond just weights, neurons, networks, etc.

Another explanation for why such an account of the internal reasoning of ANNs is theoretically possible finds itself in the very nature of a pattern. Contrasted with randomness the problem becomes quite clear. As Dennett puts it:

> [something] is random if and only if the information required to describe … the series accurately is incompressible … [contrarily, something] is not random - has a pattern - if and only if there is some more efficient way of describing it … [Therefore,] [a]

> pattern exists in some data … if there is a description of the data that is more efficient (Dennett, 1991, pp. 32–34).

While Dennett may be attempting to use his idea of 'real patterns' to support the position that one can be a realist about beliefs (Dennett, 1991, pp. 27–28), his analysis remains relevant for patterns in general. In this regard, 'real patterns' (like the kind ANNs seem to recognize) not only serve as compressed, more efficient descriptions, but, when applied to real-world data they also stand in place of real forces or phenomena. Meaning that the patterns they pick out "tend to serve [as] perspicuous representations of real forces" (Dennett, 1991, p. 29); i.e. forces such as gravity. Of course, there is variation in how efficient a given pattern is at eliminating noise and describing that which it stands in place of; but the scope of such differences does not change the core of what a real pattern is. Patterns stand in place of real forces or phenomena as simpler, more efficient descriptions of them. Think of them as compressed representations of complex real phenomena. Dennett best explains this feature of patterns with his Game of Life (GoL) example. GoL is played on a two-dimensional grid made up of cells which can either be on or off. Dennett explains the basic physics of GoL below:

> [e]ach cell, in order to determine what to do in the next instant, counts how many of its eight neighbors is ON at the present instant. If the answer is exactly two, the cell stays in its present state (ON or OFF) in the next instant. If the answer is exactly three, the cell is ON in the next instant whatever its current state. Under all other conditions the cell is OFF (Dennett, 1991, p. 37)

At the 'physical level' of GoL, there is no motion, just dots flashing in and out of existence based on the above rules. At the design level this becomes contested though. For, there are configurations of dots which 'travel' infinitely unless they encounter other dots which destroy them (Dennett, 1991, p. 39). Such configurations are known as gliders and there are other configurations like the loaf, eater, blinker, etc. Now, when an eater encounters a glider it will consume the glider entirely and return to its original configuration (Dennett, 1991, p. 40). We could choose to describe the minutiae of this interaction in its entirety at the physical level, but we could also move to the "design level, adopt its ontology, and proceed to predict - sketchily and riskily – the behavior of larger configurations or systems of configurations, without bothering to compute the physical level" (Dennett, 1991, p. 40), which is a far simpler, far more compressed description of what is going on. Instead of describing the detail of every flashing

dot, I can just say that an eater consumed a glider. Figures of a glider and a glider being consumed by an eater can be seen below.

Figure 10: *The configuration of blinking dots called a 'glider' (Dennett, 1991, pp. 39).*

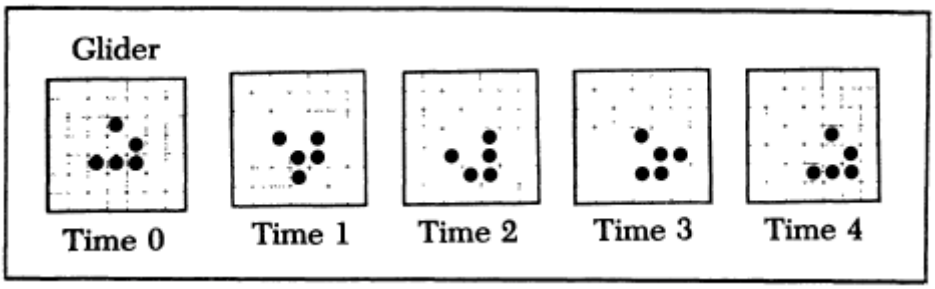
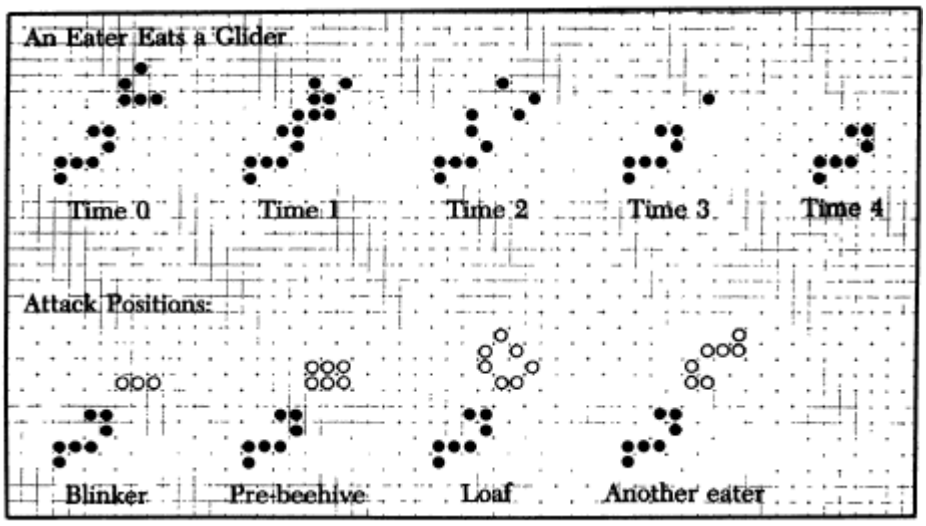


Figure 11: *An eater encountering various cellular automata (Dennett, 1991, pp. 40).*



Dennett does not stop here though. If in GoL a computer was constructed out of cellular automata like gliders and eaters (Dennett, 1991, p. 40), and it was programmed to play chess (Dennett, 1991, p. 41), then we have an even more extreme example to explain what a pattern is. Since, if we attempt to give a complete description of the chess playing computer at the physical level of the GoL, estimates place this structure at approximately 10^{13} dots in size, which would be a computationally intensive description to generate at the physical level. However, if we simply move up to

> an ontology of chess-board positions, possible chess moves, and the grounds for evaluating them; then, adopting the intentional stance toward the configuration, one can predict its future as a chess player performing intentional actions-making chess moves and trying to achieve checkmate (Dennett, 1991, p. 41).

Seen at this level of abstraction, the nature of a pattern is even more clear. When we understand that the configuration of dots represents a pattern of chess playing behaviour our descriptions of it become far simpler and more compressed representations of the forces at play. Hence, this explanation shows why the internal reasoning of ANNs should (in theory) be explainable beyond just weights, neurons, networks, etc. - the mathematical compressions of ANNs are merely outputs that stand in place of things in the real world. The output of an ANN is just a mathematical compression which represents one phenomenon whereas the chess-playing behaviour would be a linguistic compression which represents another[45]. Yet, core to both kinds of compressions is their predictive power. These 'real patterns' tend to predict future outcomes accurately regardless of whether they take the form of an ontology of chess-board positions or an ANN with a mathematical compression of your social media data. This is not to deny that the spatial reasoning of ANNs is difficult to translate into human interpretable labels. The ongoing epistemic opacity does at least indicate that the trouble we are experiencing in our attempts to translate the mathematical functions into interpretable labels is non-trivial. But, this difficulty alone does not entail that such translations are impossible either. Dennett's analysis shows why, at least theoretically, translations of these patterns should be possible. If the ANN has found a 'real pattern' then it has identified a simpler, more compressed description of a real-world force or phenomena[46]. It has just represented this compression spatially and then translated those relations into mathematical terms. There is little to suggest in the Hasson et al. account that we could not identify the same 'real pattern' with a linguistic compression or some other human interpretable label. This is because at the fundamental level, the output of an ANN is just a mathematical compression of a real-world force or phenomena. Consequently, unless there is evidence to suggest that translating mathematical compressions into linguistic compressions is impossible, we can respond to the Hasson et al. position on epistemic opacity by pointing out that patterns are just compressed representations which stand in place of real forces or phenomena, and that such compressions could be represented with a variety of techniques, some of which could consist in human interpretable rules and labels.

While I am not convinced by the Hasson et al. position towards epistemic opacity, their analysis of the interpolative processing goes a long way to explain, not the internal reasoning itself, but

[45] Although some recent literature argues that linguistic compressions are actually a sub-species of mathematical compressions (Nefdt, 2023).

[46] A pre-print by Lederman and Mahowald supports this by applying Dennett's stance more specifically to ANNs. They argue that, at least according to Dennett's intentionalist stance, ANNs do understand (Lederman & Mahowald, 2024)

why ANNs are epistemically challenging in the first place. Like brains, the kind of interpolation which ANNs use to predict and generalize requires increased parameterisation (bigger networks) and denser, more varied sampling (bigger data). Hence, combining bigger networks and bigger datasets with the already epistemically challenging FAR processing only makes them even more difficult to understand by adding increased scale to the problem. Yet, as Hasson et al. points out, the increased size of the network and input seems to be remarkably accurate of the brains interpolation (Hasson et al., 2020, p. 428). In this way, ANNs perform their tasks by compressing data and then making predictions based on their compressions; i.e. they identify real patterns. Brains, at least in conformity with the intentional stance, do the same. Additionally, according to Hasson et al., the way both networks compress and identify patterns also happens in the same way. Consequently, understanding the internal reasoning of ANNs is epistemically challenging for the same reasons as brains – we founded them using processing principles inspired by the brain and our attempts to engineer them converged on the same solutions which evolution had already engineered in the brain. Alone each factor could be coincidence but together they indicate that epistemic opacity is better explained as an emergent feature of these complex systems rather than a bug in their design.

Here I must make an earlier caveat crystal clear. Much of the current work on ANNs is concerned with the proposed pinnacle of their emergent properties; those things we call consciousness, thought, intentionality and mind. This project and the emergent properties it concerns itself with only consider the domains ANNs show competence in at the present. Which is precisely why I have limited the comparison between brains and ANNs to that of a mechanistic abstraction of the distributed spatial processing present in both networks – a claim which is sensitive to the gaps in our knowledge. For, it may well come to pass that we understand the distributed, spatial reasoning of ANNs as but one of the components of intelligence responsible for all those interesting higher-level properties. Of course, the opposite may come to pass too, where such emergent properties do arise in ANNs. However, the jury is still out on whether spatial information processing systems, especially those constituted of silicon and metal, will be capable of all the same properties as brains (Searle, 1980). Perhaps some of these ultimate cases of emergence are not substrate independent, or they are reliant on some other kind of brain processing yet unfamiliar to us, like quantum computing (Penrose, 1989, 1994).

3.4 A final objection

Objections may arise based on the functionalist links between these emergent phenomena; rejecting the similarities by accepting that they may be functionally equivalent but arguing that functional equivalency does not mean they follow all (or even enough) of the same smaller processes to achieve the larger function. This is commonly known as the inverted qualia argument, though it can extend beyond that domain. Famously this objection can be traced back to Locke:

> though one man's idea of blue should be different from another's. Neither would it carry any imputation of falsehood to our simple ideas, if by the different structure of our organs it were so ordered, that the same object should produce in several men's minds different ideas at the same time; v.g. if the idea that a violet produced in one man's mind by his eyes were the same that a marigold produced in another man's, and vice versa. For, since this could never be known, because one man's mind could not pass into another man's body, to perceive what appearances were produced by those organs; neither the ideas hereby, nor the names, would be at all confounded, or any falsehood be in either. For all things that had the texture of a violet, producing constantly the idea that he called blue, and those which had the texture of a marigold, producing constantly the idea which he as constantly called yellow, whatever those appearances were in his mind; he would be able as regularly to distinguish things for his use by those appearances, and understand and signify those distinctions marked by the name blue and yellow, as if the appearances or ideas in his mind received from those two flowers were exactly the same with the ideas in other men's minds (Locke, 1690, pp. 507 & 508).

At the heart of this problem is the idea that functional equivalency does not necessitate type identity. Functional equivalency can be achieved in a variety of ways as is expressed by Locke's example. Simply because we agree that a given colour is yellow, blue, green, etc. does not mean that we are seeing the same colour. Though, as is often the case with famous primary sources, important pieces of context are left out. Locke concludes by pointing out that, while the above is a possibility,

> I am nevertheless very apt to think that the sensible ideas produced by any object in different men's minds, are most commonly very near and undiscernibly alike (Locke, 1690, p. 508).

This seems a reasonable position to hold, especially if we are to apply Putnam's no miracle argument here too. Though, the general criticism about functionalism remains. For, we can both still call something violet while experiencing it in a different way. Applied to my argument, detractors may say that even if they granted that these systems are functionally equivalent, this would not necessitate that these systems are type identical or share functional equivalency for the same reasons.

This objection is challenging to respond to. The reality is that even if each component were functionally equivalent in brains and ANNs, it would not necessarily entail type identity in the system. However, while the evidence may not be able to support a type identity claim, that both employ interpolative processing approaches on top of all the similarities discussed in chapter 2 does give strong evidence in favour of these systems using the same strategy to generalize and make predictions. Accordingly, the objection is fair, but what it achieves is less than it appeared at first. Although it highlights that we do not have adequate knowledge on whether these systems are performing the same functions for the same reasons, the objection itself has little to say about what is actually the case, and more to say about what is possibly the case. Therefore, while a type identity claim is not supported, it is not ruled out either. In time we may come to find that functional equivalency is achieved via a different mechanism, but we may also come to find that it is achieved via a shared one. Consequently, since my argument does not rely on a type identity claim, and instead relies on a claim that ANNs are a mechanistic abstraction of the spatial processing principles of brains, I have merely given reasons to suspect that the same system configuration which generates incredible pattern recognition powers, also generates significant epistemic challenges.

Conclusion

My aim with this research is to show that there are good reasons to recouple brains and ANNs, and that this recoupling can reframe many of the problems which the pure engineering approach struggles to address. Of special significance here is that recoupling epistemic opacity with brain science allows the black box problem in ANNs to be understood as an emergent feature, not an epistemological bug.

The first chapter provides a historical link between brains and ANNs; tracing modern ANNs back to their connectionist origins as a brain model. This does not allow for brains and ANNs to be recoupled, however. For, they may have begun as a brain model, but modern ANNs have advanced significantly since Rosenblatt's day. The criticisms of early connectionist networks as well as the discovery of efficient mathematical techniques sparked revisions in methodology which would eventually push the entire field towards the engineering paradigm. Consequently, the history of ANNs shows a link with brains, but the transformation of the field during the revisionist and engineering epochs prevents a recoupling based on this history alone.

Subsequently, the second chapter explores the link between brains and ANNs - using Marr's three levels to assess just how much similarity remains between the two. At the first level, hardware implementation, there exist a variety of similarities between brains and modern ANNs, which I argue produces the same broader system structure. At Marr's second level, representation and algorithm, the shared system structure scales up to similarity in the manner both networks represent the transformation of information from input to output – since both use activation patterns for their transformations. Additionally, these activation patterns also involve the same underlying principle, using spatial relations to map patterns of statistical dependency. At the third and final level, computational theory, the computational goals of brains and ANNs share further similarity. Both kinds of networks have computational goals, and the kind of computational goals which they perform well at overlap significantly. Thus, both networks share domains of functional competency in areas like language processing, visual processing, etc. Next, to explore what kind of comparison can be justified, I extend my discussion to the East Gate Centre, a building which employs the ventilation principles of termite mounds. Concluding that, much like the East Gate Centre can be compared to termite mounds as an abstraction of their ventilation principles, ANNs can be compared to brains as a mechanistic abstraction of the brains spatial processing principles. And it is this kind of comparison which serves as the basis for why ANNs should be recoupled with brains. Yet, the

recoupling is not founded upon a claim about type identity, nor do I argue that brains and ANNs are the same. Among other limitations, understanding ANNs as a mechanistic abstraction of the brains spatial processing only provides the basis for a limited recoupling appropriate to current evidence.

Having established grounds for recoupling, the third chapter explores what the consequences of this recoupling would be for a 'problem' like epistemic opacity. In that regard, brain science offers a reframing which allows epistemic opacity to be understood as an emergent feature, not an epistemological bug. For, when we compare the criteria for emergence against epistemic opacity, the congruency between the two networks suggests that our difficulty grasping the internal reasoning of ANNs may originate from the same difficulty we have with brains. More specifically, my claim is that the epistemic difficulty comes from the shared FAR processing both networks engage in. Though, again, my claims must be tempered by current evidence; and, since our recoupling is based on ANNs as a mechanistic abstraction of the brains spatial processing, I have little to say about higher level emergent properties like consciousness, intentionality, mind, etc. Instead, my point about emergence is limited to those shared domains of functional competency - visual processing, language processing, etc. Finally, as a response to those who would object that the spatial processing of brains and ANNs is not alike enough to be the cause of shared epistemic properties like emergence, a recent paper by Hasson et al. (2020) gives evidence to the contrary. Since, they point out that brains and ANNs both process information via interpolation, and that ANN engineering has even converged on some of the same solutions as brain evolution. However, there are still some objections (like the inverted qualia objection) which remain.

Consequently, I hope I have provided good reasons to support a potential recoupling between brains and ANNs, and not just because of the interesting conclusions which can be drawn from the connection. Rather, that I have shown the similarities which remain between brains and ANNs are, on their own merit, strong enough. It is therefore an additional advantage of this argument that the subsequent recoupling of brains and ANNs has the potential to yield important insight which the engineering paradigm has thus far been unable to grant – like in the case of reframing epistemic opacity as an emergent feature.

REFERENCES

Arkoudas, K., & Bringsjord, S. (2014). Philosophical Foundations. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 34–63). Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-artificial-intelligence/philosophical-foundations/5C3626F0F8F3A9E4D5148A8DAAB908B1

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, *513*(5), 532–541. https://doi.org/10.1002/cne.21974

Babcock, N. S., Montes-Cabrera, G., Oberhofer, K. E., Chergui, M., Celardo, G. L., & Kurian, P. (2024). Ultraviolet Superradiance from Mega-Networks of Tryptophan in Biological Architectures. *The Journal of Physical Chemistry B*, *128*(17), 4035–4046. https://doi.org/10.1021/acs.jpcb.3c07936

Bauerle, A., Jönsson, D., & Ropinski, T. (2022). *Neural Activation Patterns (NAPs): Visual Explainability of Learned Concepts*. https://doi.org/10.48550/arXiv.2206.10611

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, *27*(11), 987–989. https://doi.org/10.1016/j.tics.2023.08.006

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in Philosophy of Science*, *9*, 261–325.

Boden, M. A. (2016). *AI: Its Nature and Future*. Oxford University Press.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, *195*(12), 5339–5372. https://doi.org/10.1007/s11229-018-01949-1

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), 1–19. https://doi.org/10.1111/phc3.12625

Cain, M. J. (2016). *The philosophy of cognitive science*. Polity; WorldCat.org.

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, *14*, 7–32.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, *134*(1), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Cao, R., & Yamins, D. L. K. (2021). Explanatory models in neuroscience: Part 1—Taking mechanistic abstraction seriously. *ArXiv*, *abs/2104.01490*. https://api.semanticscholar.org/CorpusID:233025412

Cappelen, H., & Dever, J. (2022). AI with Alien Content and Alien Metasemantics. In *Oxford Handbook of Applied Philosophy of Language (forthcoming)* (In Ernest Lepore (ed.), p. 20). Oxford University Press. https://philpapers.org/archive/CAPAWA.pdf

Castillo, E. F., Cobo, A., Gutiérrez, J. M., & Pruneda, R. E. (1999). Functional networks with applications: A neural-based paradigm. *The Springer International Series in Engineering and Computer Science*, *473*, 309.

Cetinic, E., & She, J. (2022). Understanding and Creating Art with AI: Review and Outlook. *ACM Trans. Multimedia Comput. Commun. Appl.*, *18*(2). https://doi.org/10.1145/3475799

Chalmers, D. J. (2006). Strong and weak emergence. In P. Clayton & P. Davies (Eds.), *The re-emergence of emergence: The emergentist hypothesis from science to religion* (pp. 244–254). Oxford University Press.

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton.

Chaves, R., & Richter, S. (2021). Look at that! BERT can easily be distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics*, *2021*, 28–38.

Chomsky, N. (1957). *Syntactic Structures* (1st ed.). De Gruyter Mouton. https://doi.org/doi:10.1515/9783112316009

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (50th ed.). The MIT Press; JSTOR. http://www.jstor.org/stable/j.ctt17kk81z

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger Publishers. https://scirp.org/reference/referencespapers?referenceid=1134882#:~:text=Article%20citationsMore%3E%3E-,Chomsky%2C%20N.,New%20York%3A%20Praeger%20Publishers.

Churchland, P. (1996a). Fodor and Lepore: State space semantics and meaning holism. In *The Churchlands and their Critics* (In R. McCauley (ed.), pp. 273–277). Oxford: Blackwell.

Churchland, P. (1996b). Second reply to Fodor and Lepore. In *The Churchlands and their Critics* (In R. McCauley (ed.), pp. 278–283). Oxford: Blackwell.

Churchland, P. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy*, *95*, 5–32.

Churchland, P. M. (2012). *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. The MIT Press. https://doi.org/10.7551/mitpress/9116.001.0001

Clark, A., & Lappin, S. (2013). Complexity in Language Acquisition. *Topics in Cognitive Science*, *5*(1), 89–110. https://doi.org/10.1111/tops.12001

Dennett, D. (1991). Real Patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dreyfus, H. (1972). *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper and Row. https://philpapers.org/rec/DREWCC

Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint. In M. Negrotti (Ed.), *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action.* (2014th ed.). Oxford University Press.

Duede, E. (2023). Deep Learning Opacity in Scientific Discovery. *Philosophy of Science*, *90*(5), 1089–1099. doi:10.1017/psa.2023.8

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Fodor, J. A. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, *28*(2), 97–115. JSTOR.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, *121*(2), 256–285. https://doi.org/10.1006/inco.1995.1136

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474. https://doi.org/10.1016/S0019-9958(67)91165-5

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Hameroff, S. (1998). Quantum computation in brain microtubules? The Penrose–Hameroff 'Orch OR' model of consciousness. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *356*(1743), 1869–1896. https://doi.org/10.1098/rsta.1998.0254

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hecht-Nielsen, R. (1995). Replicator neural networks for universal optimal source coding. *Science, 269 5232,* 1860–1863.

Hinton, G. (2007). *Boltzmann Machines*. University of Toronto. https://www.cs.toronto.edu/~hinton/csc321/readings/boltz321.pdf

Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, *14*(8), 1771–1800. https://doi.org/10.1162/089976602760128018

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "Wake-Sleep" Algorithm for Unsupervised Neural Networks. *Science*, *268*(5214), 1158–1161. https://doi.org/10.1126/science.7761831

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science (New York, N.Y.)*, *313,* 504–507. https://doi.org/10.1126/science.1127647

Hinton, G., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18(7),* 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hubel, D. H., & Wiesel, T. N. (1967). Cortical and callosal connections concerned with the vertical meridian of visual fields in the cat. *Journal of Neurophysiology*, *30*(6), 1561–1573. https://doi.org/10.1152/jn.1967.30.6.1561

Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., & Gebru, T. (2023). AI Art and its Impact on Artists. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374. https://doi.org/10.1145/3600211.3604681

Jones, C., & Bergen, B. (2023). Does GPT-4 pass the Turing test? *Https://Arxiv.Org/*, 1–28. https://doi.org/10.48550/arXiv.2310.20216

Kambhatla, N., & Leen, T. K. (1997). Dimension Reduction by Local Principal Component Analysis. *Neural Computation*, *9*(7), 1493–1516. https://doi.org/10.1162/neco.1997.9.7.1493

Karakas, O. (2023). *Consumer perception: How creative is ChatGPT really?* [Masters thesis]. Fachhochschule der Wirtschaft.

Kim, J. (1989). The Myth of Nonreductive Materialism. *Proceedings and Addresses of the American Philosophical Association*, *63*(3), 31–47. JSTOR. https://doi.org/10.2307/3130081

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *95*(1/2), 3–36. JSTOR.

Koene, R. A. (2013). Uploading to Substrate-Independent Minds. In *The Transhumanist Reader* (pp. 146–156). https://doi.org/10.1002/9781118555927.ch14

Krenker, A., Bester, J., & Kos, A. (2011). Introduction to the Artificial Neural Networks. In *Artificial Neural Networks-Methodological Advances and Biomedical Applications*. https://doi.org/10.5772/15751

Ladyman, J. 1969-, & Wiesner, K. (2020). *What Is a complex system?* (1–1 online resource (182 pages)). Yale University Press; WorldCat.org. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2552239

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lederman, H., & Mahowald, K. (2024). Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs.

Transactions of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2401.04854

Leib, R. (2023). GPT–4 WHO NOW? In *Exoanthropology* (pp. 423–436). Punctum Books; JSTOR. http://www.jstor.org.ezproxy.uct.ac.za/stable/jj.1380397.60

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*(1), e99–e108. https://doi.org/10.1353/lan.2019.0015

Locke, J. (1690). *An essay concerning human understanding* (2nd (1998 edition)). Penguin Books Ltd.

Luc, P., Neverova, N., Couprie, C., Verbeek, J., & LeCun, Y. (2017). Predicting Deeper into the Future of Semantic Segmentation. *CoRR*, *abs/1703.07684*. http://arxiv.org/abs/1703.07684

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *CoRR*, *abs/1801.00631*. http://arxiv.org/abs/1801.00631

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

McCulloch, W., & Pitts, W. (1943). A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY. *BULLETIN OF MATHEMATICAL BIOPHYSICS*, *Volume 5*, 19.

Meyes, R., Lu, M., Puiseau, C. W. de, & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. *ArXiv*, *abs/1901.08644*. https://api.semanticscholar.org/CorpusID:59291899

Millière, R. (2024). Philosophy of Cognitive Science in the Age of Deep Learning. *Wiley Interdisciplinary Reviews. Cognitive Science*, e1684.

Minsky, M., & Papert, S. (1969). *Perceptrons; an introduction to computational geometry*. MIT Press; WorldCat.org.

Mollo, D., & Millière, R. (2023). The Vector Grounding Problem. *Https://Arxiv.Org/*. https://doi.org/10.48550/arXiv.2304.01481

Nefdt, R. (2023). *Language, Science, and Structure: A journey into the philosophy of linguistics*. Oxford University Press. https://philpapers.org/rec/NEFLSA

Newell, A., & Simon, H. (1956). The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Nkandu, M., & Alibaba, H. (2018). Biomimicry as an Alternative Approach to Sustainability. *Bulletin of the Polytechnic Institute of Jassy, Constructions Architecture Section*, *8*. https://doi.org/10.5923/j.arch.20180801.01

Osborne, M. (2023). Researchers Use A.I. to Decode Words From Brain Scans. *Smithsonian Magazine*, *180982097*, N/A.

Pater, J. (2018). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *October*, 39.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*(3), 241–288. https://doi.org/10.1016/0004-3702(86)90072-X

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics.* (pp. xiii, 466). Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness* (1st ed.). Oxford University Press, Inc.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*(October 2020), 12. https://doi.org/10.1016/j.cognition.2020.104365.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz.Net*, 48.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Prinz, J. J. (2005). Empiricism and State Space Semantics. In B. L. Keeley (Ed.), *Paul Churchland* (pp. 88–112). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139165280.005

Proudfoot, D., & Copeland, J. (2012). Artificial Intelligence. In E. Margolis, R. Samuels, S. P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (p. 44). Oxford University Press.

Putnam, H. (1978). *Meaning and the Moral Sciences*. Routledge.

Ritz, T., Damjanović, A., & Schulten, K. (2002). The Quantum Physics of Photosynthesis. *ChemPhysChem*, *3*(3), 243–248. https://doi.org/10.1002/1439-7641(20020315)3:3<243::AID-CPHC243>3.0.CO;2-Y

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. https://doi.org/10.1037/h0042519

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms.* Spartan Books.

Rumelhart, & Mcclelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* (Vol. 1). The MIT Press.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Selfridge, O. (1959). Pandemonium: A paradigm for learning. *In Proceedings of Symposium on the Mechanization of Thought Processes 10*, *1*, 22. National Physical Laboratory.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Linguistic and philosophical implications), Vol. 2* (pp. xvii, 611). MIT Press.

Taylor, E. (2015). An explication of emergence. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *172*(3), 653–669. JSTOR.

Tegmark, M. (2000). The Importance of Quantum Decoherence in Brain Processes. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *61*, 4194–4206. https://doi.org/10.1103/PhysRevE.61.4194

Teh, Y. W., & Hinton, G. E. (2000). Rate-coded Restricted Boltzmann Machines for Face Recognition. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems* (Vol. 13). MIT Press. https://proceedings.neurips.cc/paper_files/paper/2000/file/c366c2c97d47b02b24c3ecade4c40a01-Paper.pdf

Teh, Y. W., Welling, M., Osindero, S., & Hinton, G. E. (2003). Energy-Based Models for Sparse Overcomplete Representations. *Journal of Machine Learning Research*, *4*(null), 1235–1260.

The Science News. (1958). "Perceptron" Thinks. *The Science News-Letter*, *74*(3), 39–39. JSTOR.

Turing, A. (1936). ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM. *Proceedings of the London Mathematical Society*, 36.

Turing, A. (1948). *Intelligent Machinery* (p. 23) [Government report]. National Physical Laboratpry.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2023). *Reclaiming AI as a theoretical tool for cognitive science.* https://doi.org/10.31234

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Warner, B., & Misra, M. (1996). Understanding Neural Networks as Statistical Tools. *The American Statistician*, *50*(4), 284–293.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science. Thesis (Ph. D.). Appl. Math. Harvard University* [PhD Thesis]. Harvard.

Wimsatt, W. C. (1997). Aggregativity: Reductive Heuristics for Finding Emergence. *Philosophy of Science*, *64*, S372–S384. JSTOR.

Wittgenstein, L. (1986). *Philosophical Investigations* (G. E. M. Anscombe, Trans.; 3rd edition). Blackwell Publishers.

Wu, M. (2024). The Effects of ChatGPT on Economic Development. *Highlights in Business, Economics and Management*, *24*, 1324–1330. https://doi.org/10.54097/tymwwt88

Yan, K., & Hricko, J. (2017). Brain networks, structural realism, and local approaches to the scientific realism debate. *Studies in History and Philosophy of Science Part C: Studies in History and*

Philosophy of Biological and Biomedical Sciences, *64*, 1–10. https://doi.org/10.1016/j.shpsc.2017.05.001

Yao, K., Cao, F., Leung, Y., & Liang, J. (2021). Deep neural network compression through interpretability-based filter pruning. *Pattern Recognition*, *119*, 108056. https://doi.org/10.1016/j.patcog.2021.108056