The Free Self: What Separates Us from Machines

by

Mitchell Ross

A thesis

presented to the University of Waterloo

in fulfillment of the degree of

Master of Arts

in

Philosophy

Waterloo, Ontario, Canada, 2023

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Could a machine ever achieve consciousness? Will it ever make sense to hold a machine morally responsible? In this thesis, I argue that the architecture of SPAUN - the largest WIP functioning brain model currently in existence - makes it the most plausible contender for strong AI status, but that a hypothetically completed, future iteration of SPAUN is not guaranteed to possess qualitative experiences, consciousness, free will, or selfhood despite its biological plausibility; it therefore cannot be held morally responsible the way we are. To justify this position, I offer critiques of determinism, compatibilism, micro-functionalism, physicalism, and naturalistic accounts of the evolution of consciousness, as well as experiments in neuroscience that appear at first glance to disprove free will. In opposition to these views, I develop a novel form of dualism which posits the self as the free, non-physical, uncaused cause of its own actions, and provide arguments to justify this position. In essence, I propose - counter to Daniel Dennett - that selves are free to do otherwise (in the classical sense), that this is their purpose, and that naturalistic accounts of the existence of selfhood, consciousness, and qualitative experiences are inadequate because they lack a view to this purpose. I conclude that because SPAUN is a physically determined system, and its underlying substrate is distinct from our own, we should be wary of ascribing cognition and moral responsibility to it, since function alone does not guarantee cognition in this novel dualistic framework.

Acknowledgements

I want to thank everyone who has helped me in my intellectual journey as a master's student at the University of Waterloo. Thank you to Chris Eliasmith for supervising my efforts, despite the depth of our disagreements, and helping me understand the intricacies of SPAUN's functioning. Thank you also to my readers Nicholas Ray and Jacqueline Feke, for taking time out of their busy schedules to read and revise my work. Special thanks to Nick for helping me interpret Dennett's works and clearly establish where mine and his philosophies diverge. I also want to thank James Chow, a fellow graduate student and friend for his help refining my understanding of some of the physics discussed in this paper. Lastly, big thanks to Anny Wang, a fellow student, friend, and cognitive science enthusiast. Our debates and discussions about determinism, qualia, zombies, neuroscience, consciousness, metaphysics, and cognitive science in general provided me with ample food for thought that helped me to further refine and distinguish the perspective offered in this paper. I could not have asked for greater support.

Table of Contents

List of Figures

Introduction

Mind-brain identity theory and physicalism have propelled neuroscientific research since the 1950s. More recently, these theories have been adopted by micro-functionalist Chris Eliasmith to produce the functional and biologically realistic brain model SPAUN (Semantic Pointer Architecture Unified Network); (Eliasmith *et al.* 2012). The success of this model in simulating cognitive behaviours and displaying inklings of general intelligence appears to provide convincing evidence in favour of identity theory, physicalism, and functionalism. As a result, it calls into question the purpose of conscious selves and the moral status of strong AI, since it may then be the case that a simulated brain can accomplish everything a biological brain can, regardless of whether that simulated brain is accompanied by a self with conscious experiences. In this thesis, I consider such consequences by asking: is cognitive behaviour sufficient for the ascription of consciousness, selfhood, or free will to SPAUN? If it is, does it then make sense to ascribe rights and moral responsibility to SPAUN? If it is not, how are we to distinguish ourselves from it even though it displays cognitive behaviours? Similar questions have been at the core of cognitive science since Alan Turing introduced functionalism and the imitation game in 1950 (Turing 1950), though the general intelligence displayed by SPAUN, and its resemblance to our own neurobiology make its capacity for conscious experience and its status as strong AI the most challenging to deny within the current landscape of AI research programmes; far more challenging than the Turing machines and Von Neumann architectures that preceded it.

It is vital that I outline the distinction between strong and weak AI at this point, since debates about the plausibility of strong AI have bedeviled AI researchers, and the field of cognitive science in general since its inception in 1950s. Unlike strong AI, weak AI already exists in abundance. It is characterized by the automation or reproduction of a narrow and specific domain of cognition. For instance, an AI chess-bot can play chess, and do nothing else. Similarly, Chat Generative Pre-Trained Transformer (ChatGPT) – a large language model-based chatbot developed by OpenAI in 2022 – can produce strings of text when provided with a user prompt, and do nothing else. These, then, are examples of weak AI whose cognitive status is easily denied, considering the narrow scope of their functioning. Strong AI, on the other hand, is a hypothetical entity characterized by the total reproduction of animal cognition (typically human cognition). In essence, strong AI is the kind of AI that would render a machine cognitively and behaviorally indistinguishable from its biological counterparts. Because biological creatures display general intelligence – the ability to switch between cognitive tasks on the fly – it is typically considered to be the primary marker denoting biological cognition, so a machine which can reproduce this kind of intelligence would be a contender for strong AI status. Due to the significance of general intelligence in characterizing biological cognition, strong AI is also sometimes referred to as Artificial General Intelligence (AGI). Unlike weak AI, there are yet to be any incontestable examples of strong AI, though I argue that because of its architecture and behaviour, SPAUN is the most promising example currently in existence, though, as per the title of this thesis, I also argue that factors beyond general intelligence also characterize human cognition, which SPAUN may never possess.

On that note, some argue that we have notions of free will able to accommodate advancements like SPAUN. For instance, Daniel Dennett's compatibilist notion of free will, which Alfred Mele dubs "modest" free will (Mele 2014, 78), is supposedly compatible with our concept of moral responsibility and criminal justice. In essence, Dennett's view is that we have free will, but not in the classical, "ambitious" sense; only in the sense that our brain allows us to consider likely future outcomes and act in accordance with our desires, though any phenomenal sense we have that we are initiating our own actions is a cognitive illusion, since in truth our desires are determined by our brains. Such a notion of free will could be ascribed to SPAUN, as well as human beings, since it construes the mind as the brain, and free beings as self-controllers whose control systems are physically determined, leaving no room for the prospect that we can do otherwise (Dennett 1984). Is this conception of free will also compatible with the phenomenal features of cognition? Can it explain the purpose of the conscious self? I think not. As a result, I argue that SPAUN, despite its success, is not guaranteed to possess what Mele dubs "ambitious" free will (Mele 2014, 79), and that it is inextricably linked to conscious selves. Mele characterizes "ambitious" free will by an addition to "modest" free will – deep openness – that would allow us to come to a different decision if placed in the exact same physical circumstances (79). What I am arguing, then, is that SPAUN, despite its cognitive behaviour, may never have a self, and therefore may never be free to "have done otherwise" the way we are. This of course all depends on whether we ourselves possess such agency, which I intend to demonstrate we do.

On that basis, I further argue that Dennett's conception of free will is too modest to account for the phenomenal experience of freely deciding as a conscious self, and that this experience can only be accounted for with a more ambitious conception of free will. These arguments require a justification that challenges the physicalism underlying both Dennettian philosophy and current AI research programmes, since "ambitious" free will (or metaphysical libertarianism) assumes incompatibility between free will and the physical determinism associated with both stances. In short, to argue that a hypothetically complete and fully operational future iteration of SPAUN would not be conscious or free in the same way that we are is to argue that there are factors beyond physical brain activity that contribute to conscious experiences like decision-making.

To that end, Chapter 1 of this thesis is devoted to characterizing SPAUN and the philosophical theories that underpin its development. The primary goal of this chapter is to establish how SPAUN's architecture and the "micro-functionalist" approach to AI development present greater challenges to the biological naturalism popularized by John Searle's famous Chinese room thought experiment than previous AIs. Chapter 2 will be devoted to characterizing Dennett's compatibilism, Harris' hard determinism, and analyzing neuroscientific experiments investigating the neuronal basis of intentional actions – ultimately to show that the accounts of Dennett and Harris leave much to be desired, and that the neuroscience data on this subject is inconclusive. Chapter 3 will be devoted to developing an argument justifying the belief that a more ambitions conception of free will is needed to account for phenomenal experiences and selfhood, as well as exploring what kind of world must exist for "ambitious" free will to be possible. This will all result in a view of the self and the world which ties "ambitious" free will to the conscious self possessed by biological beings, thus denying that non-biological entities like future iterations of SPAUN should be ascribed consciousness based on cognitive behaviour.

Chapter 1

SPAUN: The New Frontier

SPAUN presents a new frontier for AI research, the strong status of which cannot be denied based on Searle's biological naturalism[1], nor his popular Chinese room argument. As the famous thought experiment goes: Searle asks us to imagine a scenario in which we are locked in a room with a set of syntactic instructions written in English, then fed sentences written in Chinese characters and expected to use the instructions to produce a meaningful response in Chinese. This scenario is essentially an analogy comparing us in the room to a CPU in a computer and is meant to demonstrate that meaningful sentences can be produced by syntactic manipulation without semantically understanding the terms of the sentence, a fact which challenges the notion that intentional states should be ascribed to computer systems based on their behaviour.

The reasons that this thought experiment cannot present a meaningful challenge to SPAUN's status as a strong AI are three-fold: (1) semantic pointer models like SPAUN are capable of replicating deep semantic processing (Eliasmith 2013, 99), the very cognitive feature which Searle's Chinese room thought experiment is typically employed to deny machines possess (Searle 1980); (2) SPAUN's architecture takes inspiration from computational neuroscience and neuroanatomy, meaning modelled brain areas are functionally and geometrically equivalent to those of real, biological human brains; and (3) the theories underpinning SPAUN's development – namely physicalism and identity theory – also underpin Searle's biological naturalism. Add to these considerations the fact that Searle does not deny the possibility of artificially reproducing consciousness on the grounds of biological naturalism so long as the artificial architectures in question are neurobiologically plausible (Searle 2007, 328). These challenges, while

[1] Biological naturalism: Searle's proposed solution to the mind-body interaction problem. Biological naturalism proposes that lower-level neurobiological processes produce all mental phenomena, and that these mental phenomena are higher-level features of the brain.

significant, are not insurmountable by biological naturalists and bio-chauvinists[2]. They could still argue, for instance, that SPAUN is a digitally simulated model running on software (Nengo) and is therefore incapable of conscious understanding in the same way that an NWP (Numerical Weather Prediction) model is incapable of getting anything wet. This argument is *prima facie* compelling, though it appeals to the engineering problem to its own detriment. Given the rate at which our technologies have advanced since the 1950s, there is reason to believe that a functional model like SPAUN could one day be realized as a physical system, so time will tell if this objection stands.

That said, the argument I intend to make does not appeal to the engineering problem. I instead argue that a completed and physically realized functional brain model like SPAUN is not guaranteed to possess qualitative experiences (a.k.a. qualia), consciousness, or free will despite its resemblance to the human brain. I am going to present a challenge to the physicalism that underlies both SPAUN's development and Searle's biological naturalism. More specifically, I am going to argue that conscious selves do not emerge solely from brain activity, but from the meeting between the physical brain and some indeterminate, potentially non-physical entity or force. This view may be interpreted as substance dualism, or neutral monism, though for the purposes of this thesis, I argue for a dualistic view since it is the more controversial among the two. Moreover, despite the resemblance of this argument to that of David Chalmers, who argues that cognitive science should view consciousness as a fundamental feature of reality (Chalmers 1997, 277) – something I agree with – I am not arguing for pan-psychism. This is because pan-psychism ascribes consciousness too liberally. Suffice it to say that anyone who would ascribe consciousness to rocks would have no problem ascribing it to AIs regardless of their biological plausibility.

Before I can do any of this, I first need to explain what SPAUN is, what it can do, and how it functions. A disclaimer is in order here: I am not a computer scientist, computational neuroscientist, or AI researcher. SPAUN is a highly complex system engineered to resemble the most complex system we know

[2] Bio-chauvinism: That biological beings have an inherent superiority over non-biological beings in terms of reproductive capabilities and intelligence. In this view, conscious experience is causally reducible to brain activity, but not ontologically reducible to physicality.

of in the animal world: the human brain. As a result, the upcoming section may be technical and difficult to understand at times, since I may not do the best job of explaining the technicalities of SPAUN's architecture. That said, I am not going to focus too heavily on the technicalities, but rather the theories underlying SPAUN's development and what the model is capable of. This is because I do not intend to make a technical argument – since that would invoke the engineering problem – but a philosophical argument. Thus, I will often defer to Eliasmith and will only give as much detail as necessary to relay SPAUN's biological plausibility.

SPAUN's Semantics

SPAUN is a WIP, large-scale, functional neural model currently comprised of 6 million spiking neurons and 8 billion connections. Its development began in 2011, and the original iteration was modeled using Nengo 1.4, a free Java package created specifically to simulate large-scale neural models like SPAUN (Eliasmith). The very first generation of Nengo (called NEsim) was developed in 2000 by Eliasmith and Anderson, and was released as a set of MatLab scripts, rather than a Java package (Bekolay *et al.* 2014). Over the years, four subsequent generations of Nengo have been released, culminating in the most recent fifth generation version dubbed Nengo 2.0, which was released as a python package in 2015 and remains under widespread use and development. Despite being originally released by Eliasmith and Anderson, it was only with the creation of the third generation that the software became powerful enough to support SPAUN's development, though simulation speeds still left much to be desired (Bekolay *et al.* 2014). When it was originally released, despite running on Nengo 2.0 and a powerful supercomputer, 1 minute of simulation time took approximately 2 hours to simulate (Eliasmith 2013, 8:40). The current iteration of SPAUN is much larger, however, and incorporates multiple GPUs to run the simulation at approximately 60% real-time.

The architecture used to construct the SPAUN model is especially interesting and relevant when it comes to challenging those who habitually employ the Chinese room argument to deny the intentional status of machines. This is because, as per its acronymous title, SPAUN is a unified network constructed by means of the semantic pointer architecture (SPA); (Eliasmith 2013). This architecture does a particularly good job of simulating deep semantic processing, or intentionality, and this is why the Chinese room argument cannot be employed to challenge the notion that SPAUN possesses intentionality in any meaningful way. This is not to say, as mentioned earlier, that there are no other avenues available by which one could argue this same point; it is merely to say that this most famous, go-to argument is no longer compelling in light of SPAUN's recent success. To demonstrate how the SPA is capable of modelling deep semantic processing, I provide a brief characterization and overview of its implementation.

First off, a disclaimer is in order: the SPA, like SPAUN itself, is a work in progress. The SPA is the operationalization of the Semantic Pointer Hypothesis which, though adequately generalizable and able to account for many cognitive behaviours, does not yet constitute a complete and unified theory of cognition. Most notably, it only characterizes mature cognitive systems, and has little to say about the evolutionary and early developmental contributions to their maturation (Eliasmith 2013, 88). Nevertheless, the implementation of the SPA is worth pursuing for its basis in theoretical neuroscience and capacity for simulating deep semantics. The Semantic Pointer Hypothesis is defined as the following: "Higher-level cognitive functions in biological systems are made possible by semantic pointers. Semantic pointers are neural representations that carry partial semantic content and are composable into the representational structures necessary to support complex cognition" (90). How, then, do semantic pointers function to capture partial semantic content?

To start with, a clarification of what makes semantic processing "deep" as opposed to "shallow" is necessary. This distinction was introduced by Allan Paivio in 1971 as an integral feature of his Dual-Coding Theory (96). This theory proposes that linguistic and perceptual information are processed through distinct channels: the linguistic using a "symbolic code," and the perceptual using an "analog code." Studies in neuroscience by Solomon and Barsalou in 2004 and Kan *et al.* in 2003 provide evidence for this theoretical distinction, since simple lexical association tasks can be completed with relative speed when compared to more difficult ones, and fMRI data shows that activation of perceptual systems in the brain only occur when considering more difficult associations (96). These distinct channels through which information is processed in the brain make for the distinction between shallow and deep semantic processing.

Shallow semantic processing is entirely linguistic and relatively fast (97). It is therefore analogous to the linguistic channel proposed by Paivio's Dual-Coding Theory. It is most easily demonstrated by our proclivity to form associations between words while oblivious to the referents of the words themselves. Reciting words that immediately come to mind in response to a prompt word is a task which involves shallow semantic processing. For example, when prompted with the word "hot," one might reply with "cold," or even "hat." While one of these replies is not like the other, neither requires that one consult an image stored in memory to make the association, despite the association between "hot" and "cold" requiring an understanding of their referents, unlike the "hat" reply. Deep semantic processing, on the other hand, does require that one consult the representations stored in memory, and is therefore analogous to the perceptual channel proposed by the Dual-Coding Theory. It is characterized by "a kind of 'simulation' of the circumstances described by the linguistic information" (97) carried by a word: its meaning. Imagining and describing a situation involving the referent of a word is a task involving deep semantic processing. This sort of processing is slower paced than shallow processing, as one is not merely activating words associated with a prompt word, they are required instead to activate a mental representation of the referent itself in working memory, including but not limited to tactile, emotional, auditory, or visual information (97).

The distinction between shallow and deep processing carries significance when considering the challenge the SPA presents to the Chinese room argument. If the SPA were only able to capture shallow semantics, for instance, then the objection that computers lack intentionality would still stand, since Searle, sitting in the room and writing Chinese characters according to syntactical instructions written in English, could still form associations between characters while remaining oblivious to their referents. In other words, he could still semantically process the characters shallowly, since shallow processing does not necessarily require that we "simulate" the referents of words or characters themselves, or that we possess a concept that carries semantic or referential content. It seems that the point of the Chinese room thought experiment is really to show that deep semantic processing is required for intentionality. Thus, because the SPA provides a clear means of not only explaining, but also implementing deep semantic processing in models like SPAUN, even Searle must admit that SPAUN has intentionality. SPAUN has an architecture that allows it to possess concepts, and therefore semantic content. It goes beyond mere manipulation of symbols based on their syntactic characteristics. Now, I must explain what semantic pointers are, and how the SPA is able to capture both shallow and deep semantics.

A semantic pointer is a high-dimensional vector in a vector space (90). In simpler terms, your typical vector is a mathematical construct signified by an arrow which is characterized as having a length and a direction. A high-dimensional vector is defined relative to a large set of orthonormal[3] vectors sharing the same point of origin. More specifically, a single high-dimensional vector is a weighted sum of the set of orthonormal vectors. In fact, this large set of vectors defines an entire family of vectors, or a vector space. This vector space is employed to describe the mathematical state space in neurobiological systems (57). In other words, one high-dimensional vector can represent the information encoded by a whole population of neurons representing a single concept. Thus, the individual orthonormal vectors of which a high-dimensional vector is comprised can represent either a single example associated with a given concept or a single feature, whereas the high-dimensional vector itself represents the whole concept (91). It is also important to note that semantically similar examples are clustered in relatively close proximity when compared to dissimilar ones (91). These proximity relations are often thought to mimic the conceptual proximity of the exemplars[4] we ourselves use when categorizing objects. Thus, the mean distribution of

[3] Orthonormal: Both orthogonal and normalized. For vectors to be orthogonally related is for them to be perpendicular, forming right angles between one another. For vectors to be normalized is for them to have the same length.
[4] Exemplars: A resemblance theory of concepts and categorization. It posits that we categorize new objects by comparing them to multiple similar members (a.k.a. exemplars) of a category already stored in memory.

orthonormal vectors for one high-dimensional vector is analogous to the prototype[5] of the concept (108, 368). In this way, high-dimensional vectors capture empirically validated theories of categorization.

The fact that the vectors of which the high-dimensional vector is comprised are orthogonal requires that the dimensionality of the vector space in which the high-dimensional vectors are located is similarly high-dimensional, and in SPAUN the vector space is able to accommodate vectors of up to 500 dimensions (158). The reason for individual vectors being orthogonally related and for the vector space being high-dimensional is to keep the vector space uncluttered. Such a cluttered space would lead to confusion between neighbouring concepts given the presence of noise and uncertainty within the system (91).

From this brief overview it should be clear how the SPA can, in theory, use high-dimensional vectors to represent a concept encoded in a population of neurons, though some further clarification is needed to demonstrate how these vector space representations (VSRs) can be used to model deep semantics. To explain this, I first need to clarify the meaning of the term "pointer." A pointer is essentially a set of numbers that acts as an address for information stored in memory, typically in a digital computer (93). One of the nice things about pointers that contributes to their ability to model both deep and shallow semantics is that they can be employed by a system to indirectly make use of information stored at an address (93). On top of this, the pointer can also be "dereferenced" to directly access the content stored at the address, and this process is integral for capturing deep semantics (101), since "dereferencing," in tandem with the realistic modelling of perceptual brain areas (most notably the visual cortex in the case of SPAUN), provides the detailed perceptual information required for the kind of "simulating" that characterize deep semantic processing (109).

Thus, addresses are also reminiscent of symbols, in the sense that they gain their computational utility from the arbitrary relationship they maintain with the content of the information stored within them

[5] Prototype: A resemblance theory of concepts and categorization. It posits that we categorize new objects by comparing them to the most typical, abstracted member of a category (a.k.a. the prototype) already stored in memory.

(93). In other words, we use the symbol "cat" to refer to the domesticated felines we all know and love, though the symbol "cat" itself is only related to this referent arbitrarily, since we could have, if we so wished, instead used the symbol "cat" to refer to dogs, and the symbol "dogs" to refer to cats. For us, these arbitrary relations between symbols and referents are learned over the course of our individual development and evolution as a species. In SPAUN's case, these relations are not arbitrary, since the pointers are derived from the object representations, a fact further demonstrating the SPAUN's limitations in terms of characterizing only mature cognitive systems, though the modelers have been careful to base their design decisions on neuroscience data gathered from studying neurobiological systems (6).

It may seem as though I have described two separate entities over the course of this section: addresses and high-dimensional vectors. I must clarify then that a semantic pointer is truly both, and this apparent difference only results from describing the same mathematical entity from different levels of abstraction (95). From the computational standpoint, a semantic pointer is a vector; from a functional standpoint, a semantic pointer is an address (94). Thus far, I have only described how the SPA can be employed to model perceptual areas of the brain, but as a final note regarding the virtues of the SPA, this architecture is also robust enough to model motor areas of the brain (112). In fact, it can model every region of the brain in theory. Unfortunately, getting into the details about how the SPA captures motor functions will take me too far afield, but this fact will become evident when I describe SPAUN's structure and what it can do, a task to which I now turn.

SPAUN's Structure

As previously mentioned, SPAUN is a functional brain model which demonstrates an unprecedented level of general intelligence, making it by far the most convincing contender for strong AI status to date. This is despite the fact it is currently comprised of 6 million neurons, which in comparison to our own brains comprised of approximately 86 billion neurons, is a relatively small number. Nevertheless, Chris Eliasmith and his team have succeeded in functionally modeling multiple brain areas with these 6 million neurons to great effect. An overview of SPAUN's neuroanatomical structure and cognitive functions is required to demonstrate how similar SPAUN's behaviour is to our own. For clarification, I am going to be describing the makeup and functions of BioSPAUN: a more recent, augmented, and more biologically realistic iteration of SPAUN than that described in Eliasmith's *et al.* 2012 paper. For the purposes of this paper, however, I use the term "SPAUN" to refer to this more recent 2016 iteration as well.

Neuroanatomically speaking, SPAUN is currently comprised of a multi-layered visual cortex (V1/V2/V4), anterior-inferotemporal and inferotemporal cortices (AIT/IT), dorso-lateral and ventro-lateral prefrontal cortices (DLPFC/VLPFC), an orbito-frontal cortex (OFC), a posterior-parietal cortex (PPC), a primary motor cortex (M1), a supplementary motor area (SMA), a premotor cortex (PM), a ventral striatum (vSTR), a subthalamic nucleus (STN), globus pallidus externus and internus (GPe/GPi), substantia nigra pars compacta and reticulata (SNc/SNr), a ventral tegmental area (VTA), and a thalamus (Eliasmith 2016, 5). The thalamus itself is also comprised of subthalamic brain areas, that I will further detail at a later point. Figure 1, reproduced from Eliasmith *et al.* 2012, demonstrates the neuroanatomical and functional structure of the model (see page 15).

The functioning of the visual cortex is crucial for processing incoming information from SPAUN's simulated eye, and the functioning of the motor areas is crucial for controlling SPAUN's physically modelled arm (not shown in Figure 1), which has mass, inertia, length, etc. (Eliasmith *et al.* 2012, 1202). Keep in mind that by "physically modelled" I do not mean to say that SPAUN is equipped with a physical, robotic arm. This is only to say that the computer modelling of its simulated arm takes inspiration from biology and physics in similar fashion to the rest of its components. Chris Eliasmith and his team are currently working on integrating a real, physical robotic arm into SPAUN's makeup, but computing limitations have thus far made this integration impossible, since the simulation time and real-time are incongruent.

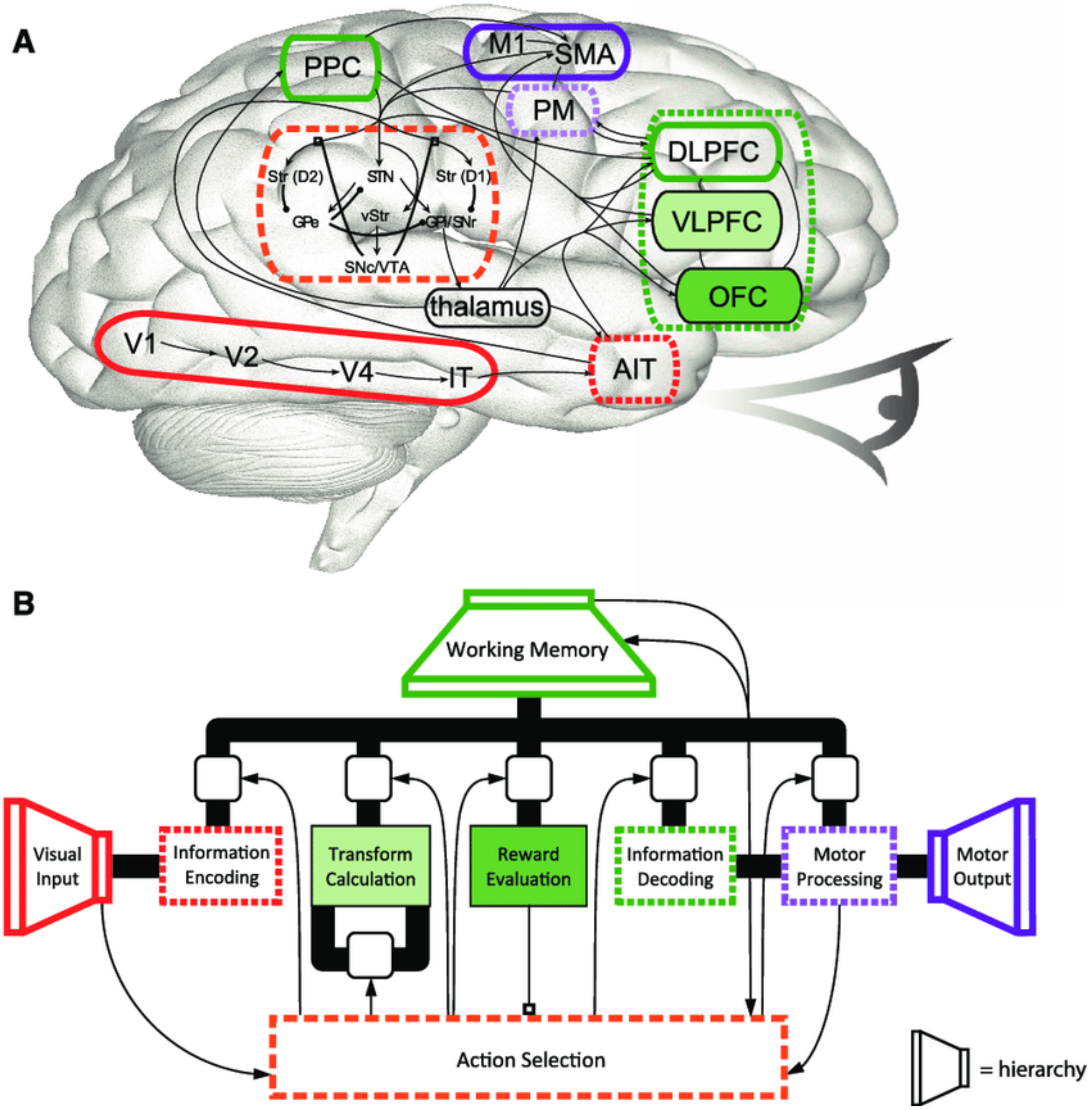

Figure 1: Anatomical and functional architecture of SPAUN. (A) The anatomical architecture of SPAUN shows the major brain structures included in the model and their connectivity. Lines terminating in circles indicate GABAergic connections. Lines terminating in open squares indicate modulatory dopaminergic connections. Box styles and colors indicate the relationship with the functional architecture in (B). PPC, posterior parietal cortex; M1, primary motor cortex; SMA, supplementary motor area; PM, premotor cortex; VLPFC, ventrolateral pre- frontal cortex; OFC, orbitofrontal cortex; AIT, anterior inferior temporal cortex; Str, striatum; vStr, ventral striatum; STN, subthalamic nucleus; GPe, globus pallidus externus; GPi, globus pallidus internus; SNr, sub- stantia nigra pars reticulata; SNc, substantia nigra pars compacta; VTA, ventral tegmental area; V2, secondary visual cortex; V4, extrastriate visual cortex. (B) The functional architecture of SPAUN. Thick black lines indicate communication between elements of the cortex; thin lines indicate communication between the action- selection mechanism (basal ganglia) and the cortex. Boxes with rounded edges indicate that the action- selection mechanism can use activity changes to manipulate the flow of information into a subsystem. The open-square end of the line connecting reward evaluation and action selection denotes that this connection modulates connection weights (reproduced from Eliasmith et al. 2012, 1203).

Besides the perceptual and motor areas, the “cortex-basal ganglia-thalamus loop,” which involves the frontal cortices, the vSTR, the STN, the GPe/GPi, the SNc/SNr, and the thalamus, plays a crucial role in SPAUN’s action selection for both fixed and flexible sequences of action. The frontal cortices store and manipulate representations, the basal ganglia[6] maps the current brain states to future brain states by selecting appropriate actions, and the thalamus monitors the rest of the system in real-time (Eliasmith 2013, 185). The thalamus is particularly integral to this whole procedure, as it can “generate gating signals that control the flow of information between visual input and working memory” (201), thereby cutting down on errors in SPAUN’s visual attention by providing it the means to ignore irrelevant visual information, which is necessary for capturing more flexible sequences of action (201). It is necessary to clarify at this point that SPAUN’s simulated eye is fixed (this is not to say its eye is physical), restricting its control over its visual input considerably (Eliasmith *et al.* 2012, 1205). Nevertheless, the selective routing of visual information still needs to be accounted for, as it is by SPAUN’s action selection mechanism, namely, the “cortex-basal ganglia-thalamus loop” (Eliasmith 2013, 185). Implementation of this loop for action selection is well justified by neuroscientific research, and thus provides a means of modelling routing and action selection procedures in a biologically plausible manner (201).

On top of this, four different neurotransmitters (GABA, AMPA, NMDA, and Dopamine) are included in the SPAUN model, as their known synaptic effects can all be simulated in the SPA (Eliasmith 2016, 3). These neurotransmitters play specific and integral roles in different procedures within the brain. GABA and AMPA play crucial roles in inhibition and excitation respectively, both of which are necessary for controlling the output of motor signals (Eliasmith 2013, 181). This makes their modelling integral for the implementation of ARC for routing information and the implementation of the “cortex-basal ganglia-thalamus loop” for action selection (181). More specifically, the basal ganglia is mostly comprised of GABAergic neurons (Eliasmith *et al*. 2012, 1202), while the striatum is comprised of dopaminergic neurons

[6] The basal ganglia: A group of subcortical nuclei including the striatum (STR), the globus pallidus (GP), the ventral pallidum (VP), the substantia nigra (SN), and the subthalamic nucleus (STN).

(1202) that play a role in modulating the routing of information and the “cortex-basal ganglia-thalamus loop” (Eliasmith 2013, 243) since, more generally, dopamine acts as a reward mechanism in the brain, and is therefore necessary for learning and conditioning. Lastly, NMDA and AMPA synaptic connections are largely restricted to the cortex and cerebellum (Eliasmith *et al.* 2012, 1202).

SPAUN’s Abilities

Having characterized the SPA and SPAUN’s neuroanatomical structure, I now turn to the task of relaying its various functions. Currently, SPAUN can perform 8 different cognitive tasks, without the need for intervention by modelers when switching between tasks (1202). These tasks include copy drawing, image recognition, a three-armed bandit/reinforcement learning (RL) task, a list reproduction/remembering (serial working memory) task, counting, question answering, rapid variable creation, and fluid reasoning. I emphasize here that modelers need not intervene when SPAUN switches tasks because it evinces the claim that SPAUN displays an unprecedented level of general intelligence. This general intelligence is unique to SPAUN, when compared to other large cognitive models. Even though some concurrent AIs demonstrate superhuman intelligence – AI chess bots for instance – these systems can only perform one task: playing chess, in this case. Without significant reprogramming, your typical chess bot would be incapable of playing any other game, even checkers. This ability to switch between tasks on-the-fly characterizes the general intelligence we possess as biological creatures. In addition, SPAUN is doubly unique since, with regards to brain models in general, it is currently the largest functional model of the human brain (1202). In comparison to its superintelligent counterparts, SPAUN displays no capacity for superintelligence on any of the cognitive tasks it performs, since this would detract from is biological plausibility. On the contrary, the purpose of the SPAUN project was never to make an AI, but to understand how the human brain works, thus it makes the same kinds of errors we do when performing these tasks, and its overall performance resembles that of the average adult human (1203).

Because SPAUN exists in a simulated environment, it is important that I provide some clarification about how it performs these tasks within its unique environment. Being a work in progress, SPAUN is only capable of dealing with digits between 0 and 9, and a handful of other symbols, namely: question marks which prompt it for replies; arrows which denote the beginning and end of lists, or the termination of a row in a matrix; and a couple of letters denoting the kind of question it is being asked. Most often, these symbols are displayed in a typed case, though handwritten symbols are also displayed as input for some tasks (i.e., Copy Drawing). All information is displayed on a simulated screen located directly in front of SPAUN's eye in the form of 28x28 pixel .jpeg images (1202). As previously mentioned, SPAUN is also equipped with a digitally reproduced arm which has 2 joints and can be used to write responses on a simulated notepad directly beneath the screen. Figure 2, reproduced from Eliasmith *et al.* 2012, displays the layout of SPAUN's digital environment (see page 19).

Figure 2 also depicts the thought bubbles which display raw neuronal activity in specific regions of SPAUN's brain besides the Str and GPi. In addition, they display the decoding of the information encoded by its spiking neuronal activity for the purposes of verifying that the model is accurately representing incoming visual information. For example, the top-right thought bubble displays the digits 4326, indicating that SPAUN is currently holding this list of digits in working memory, since the thought bubble in question displays activity occurring in the DLPFC. The bottom-left thought bubble displays a decoded question mark represented in IT, indicating that SPAUN recognizes it is being shown a question mark on-screen. I will save a more detailed description of the encoding and decoding procedures for the upcoming section. For now, I instead turn to the task of laying out and describing 8 cognitive tasks and SPAUN's performance on them in detail. It is important to note that the most recent 2018 iteration of SPAUN performs 12 cognitive tasks, though this is a recent development with no documentation yet available. I therefore restrict this section to outlining tasks performed by the 2016 model. For transparency's sake, I also provide a link to online videos showing SPAUN's performance on these tasks (https://xchoo.github.io/spaun2.0/videos.html):

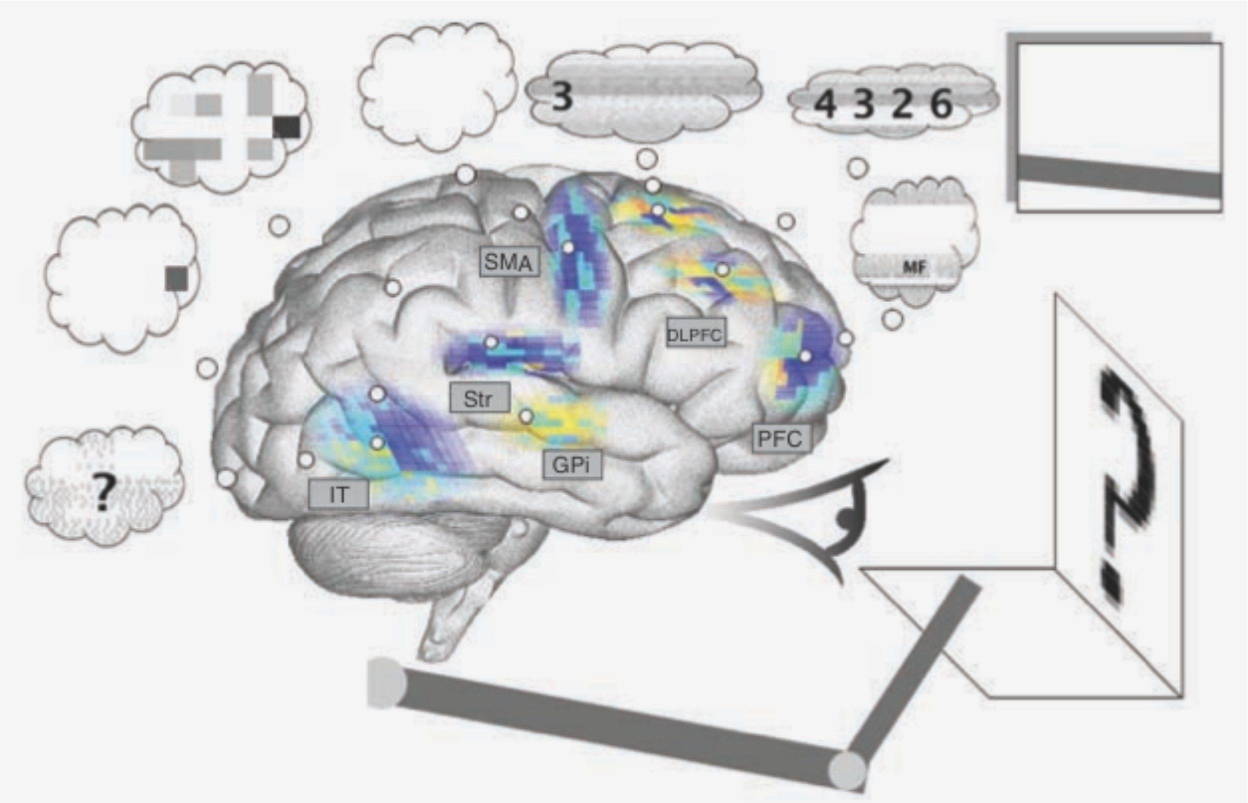

Figure 2: A screen capture from the simulation movie of the serial working memory task. The input image is on the right, the output is drawn on the surface below the arm. Spatially organized (neurons with similar tuning are near one another), low-pass–filtered neuron activity is approximately mapped to relevant cortical areas and shown in color (red is high activity, blue is low). Thought bubbles show spike trains, and the results of decoding those spikes are in the overlaid text. For Str, the thought bubble shows decoded utilities of possible actions, and in GPi the selected action is darkest (reproduced from Eliasmith *et al.* 2012, 1204).

1. Copy Drawing: SPAUN is shown a random handwritten or typed digit which it then reproduces in the same style (1202).
2. Image Recognition: SPAUN is shown a random handwritten digit which it then reproduces in its default style (1202). SPAUN's accuracy rate in recognizing these unfamiliar hand-written digits is 94% (Eliasmith 2016, 11), approaching the ~98% accuracy of the average human (Eliasmith *et al*. 2012, 1204).
3. Three-armed Bandit Task (RL): SPAUN is prompted to guess one of three digits between 0 and 2 and is rewarded if it guesses correctly. SPAUN will change its guess until it learns what response is most highly rewarded. SPAUN will also continue to guess the same digit for which it was previously rewarded until the reward is rescinded multiple times. Reward contingencies for this

task can be changed from trial to trial (Eliasmith *et al.* 2012, 1202). Neuronally speaking, the dynamics of the firing rate changes in SPAUN's striatum during this task reflect that of the mammalian brain (Eliasmith 2016, 4).

4. List Reproduction/Remembering (Serial WM) Task: SPAUN is shown a list of arbitrary length composed of digits from 0 to 9 which it then reproduces in its default style (Eliasmith *et al.* 2012, 1202). Much like us, SPAUN makes errors in its reproduction of a list exceeding 5 digits. As with people, these errors are more likely to occur when reproducing terms in the middle of such lists (Eliasmith 2016, 4).
5. Counting: SPAUN is shown a starting value and a count value to be added to the starting value. It then counts "silently" in working memory and reproduces only the final value (Eliasmith *et al.* 2012, 1202). SPAUN's mean reaction time increases with the difference between the starting value and the final value, reflecting a feature of human cognition called Weber's Law (Eliasmith 2016, 4).
6. Question Answering: SPAUN is shown a list of digits, after which it is prompted to answer a P question or a K question. A P question asks which digit occupied a specific place in the list, whereas a K question asks which place in the list was held by a specific kind of digit (e.g., even/odd) (Eliasmith *et al.* 2012, 1202).
7. Rapid variable creation: SPAUN is shown examples of syntactic input/output patterns (e.g., input: 0014, output: 14). It is then shown a novel input pattern (e.g., 0074) and is prompted to produce the novel output (e.g., 74) (1202).
8. Fluid reasoning: SPAUN performs an induction problem akin to the Raven's Progressive Matrices[7] (RPM) test for fluid intelligence. Its performance on these RPM tests reflects that of the average human being (Eliasmith *et al.* 2012, 1202; Eliasmith 2016, 4).

[7] Raven's Progressive Matrices: A non-verbal test typically used to measure the general intelligence and abstract reasoning that characterizes fluid intelligence. Test questions usually consist of a set of visual geometrical figures with the last figure missing (to be filled in by the testee).

As if all this were not compelling enough to convince you of its biological plausibility, SPAUN is also capable of simulating the effects of drugs on the brain. More specifically, the 2016 iteration of SPAUN has been used to simulate the cognitive behavioural effects of the drug tetrodotoxin (TTX) (Eliasmith 2016, 1). It is important to note that simulating cognitive behaviours and the effects of drugs alone is not compelling evidence of SPAUN's biological plausibility. The real evidence is rather the fact that SPAUN's brain-like architecture enables modelers to simulate processes internal to the brain that result in these behaviours and effects as well. To run this simulation, SPAUN had to be augmented from leaky integrate-and-fire (LIF) neuron models to a hybrid between LIF and compartmental neuron models. The compartmental model is simulated in NEURON 7.1 and adds 20 compartments across 4 functional neuronal areas (soma, basal dendrite, apical dendrite, and apical dendrite tuft) and 9 separate ion channels (6). These ion channels are especially important because TTX is known to block the sodium ion channels in neurons (6), and the relatively simple LIF model does not include them. A Python interface called *nengo_detailed_neurons* between Nengo and NEURON was implemented to accommodate the compartmental model (6). It is important to note that SPAUN's performance on the previously laid out cognitive tasks was not adversely affected when replacing the LIF neurons in IT and OFC with compartmental neurons (8, 11). That is, the compartmental augmentation alone (without simulating the effects of TTX) did not affect SPAUN's performance.

SPAUN's performance on the image recognition task was significantly impacted, however, by blocking 72% of the ion channels in IT. At this blockage percentage, SPAUN's classification accuracy dropped from its usual 94% to a mere 10% (6). This is because the blocking of ion channels drastically increases the noise within the system (8), as well as its root-mean-square deviation (RMS) error rate (6). SPAUN's performance on the counting and list reproduction tasks was also impacted by blocking only 20% of the ion channels in OFC. For example, on the counting task, SPAUN encodes the starting value and the count value with no issues, but then forgets the task state since its memory is interrupted by the blockage of ion channels. This results in its being unable to complete the counting task and hence producing no output

(11). A video showing this effect can be seen by following this link: <https://www.youtube.com/watch?v=FoOGqzG8_WU>. Similarly, on the list reproduction task, SPAUN also encodes the list of digits with no issues but is unable to reproduce the whole list since it forgets the task state before the task can be completed (11). A video showing this effect can be seen by following this link: <https://www.youtube.com/watch?v=kpwoBccdmd8>. These simulations of the effects of TTX show that the model can be used to develop medical tests to determine how new drugs will affect the behaviour of human patients (12).

This brief overview of SPAUN's structure and function demonstrates its biological plausibility and its status as a contender for strong AI model, despite its relatively small size compared to biological brains and the fact that the original purpose of the model was never to construct an AI at all. I have not yet touched on the philosophical underpinnings of the engineering framework implemented by its modelers, but this task will prove crucial for establishing the relatively new paradigm cognitive scientists find themselves in. The next section is devoted to characterizing this engineering framework and its philosophical underpinnings.

Neural Engineering

SPAUN embodies the Neural Engineering Framework (NEF) devised by Chris Eliasmith and Charles Anderson (Eliasmith & Anderson 2003). The NEF is comprised of three core principles outlined by Eliasmith in his 2013 book *How to Build a Brain*:

> 1. Neural representations are defined by the combination of nonlinear encoding (exemplified by neuron tuning curves, and neural spiking) and weighted linear decoding (over populations of neurons and over time).
>
> 2. Transformations of neural representations are functions of the variables represented by neural populations. Transformations are determined using an alternately weighted linear decoding.

> 3. Neural dynamics are characterized by considering neural representations as state variables of dynamic systems. Thus, the dynamics of neurobiological systems can be analyzed using control (or dynamics systems) theory.
>
> (Eliasmith 2013, 49)

These three principles, in tandem with a fourth overarching principle stipulating that the noisiness of neural systems must be accounted for (49), are inspired by theoretical neuroscience and experimental findings in computational neuroscience (47). Together they provide a means of simulating neural systems in a biologically realistic manner since they demonstrate how to map computations onto a substrate with many of the same constraints as biological brains. Importantly, the NEF does not describe in any way *what* neural systems and simulations are doing, only *how* they might compute over representations (48); *what* is stipulated not by the NEF, but by the modeler employing it (49). At this point, some clarification of these principles is in order, since each is based on relevant theories in philosophy of mind related to the emergence of the mind from brain activity. Keep in mind also that these principles do not necessarily function independently of one another. That is, the second principle builds on the first principle, and the third principle builds on the second and first principles, so one might also consider them to be steps in the process of characterizing the function of cognitive systems, culminating in the third principle.

The first principle provides an information-theoretical definition of neural representation hinging on the concept of a code that includes encoding and decoding procedures. This is to say information is "encoded" in the brain by mapping stimuli to spiking neurons (50). As Eliasmith points out, this is usually where the theoretical neuroscience halts in its inquiries, but it is not enough to characterize the lived mental experience, nor to verify that stimulus information is adequately represented by spiking neuron populations in SPAUN (51). This is because our lived mental experience is nothing like the experience, say, of encoded information in spiking neurons. When we see a bus, for instance, we do not recognize it as a bus because we feel that our "bus neurons" just fired. We see the bus as a bus. This is to say, that the "bus information" is represented to us internally, and qualitatively in a *decoded* manner, because "If no information about the stimulus can be extracted from the spikes of the encoding neurons, then it makes no sense to say that they represent the stimulus" (51). In other words, if the "bus information" encoded in spiking neurons did not enable us to recognize buses as buses, then it makes no sense to say we possess a representation of a bus in mind. This is a crude analogy, and I only include it to demonstrate the significant role decoding plays in *explaining* how the mind functions from an information-theoretical standpoint. As far as we know, there is no concrete evidence suggesting that neurobiological systems decode anything, though SPAUN's successes lend some credence to this idea. If such decoding procedures are necessary, and they do not occur in the brain, it could otherwise be the case that these procedures result from the interaction between the physical and the non-physical. That is all hypothetical of course, though the fact that it is questionable whether biological brains themselves need to respect these decoding procedures despite their activity being accompanied by conscious experience could be counted as a point in favour of dualist interpretations. It is important to keep in mind that both the encoding and decoding procedures are theoretical constructs which ultimately serve the purpose of verifying the representational function of the system, as I mentioned in the previous section (62).

The decoding procedure is much more complicated than the encoding procedure, and it has two aspects: the population aspect and the temporal aspect. The former accounts for the fact that information about a single stimulus is typically encoded by a population of neurons instead of just one neuron, while the latter accounts for the fact that populations of neurons respond in time to a changing stimulus (51). The details of this decoding procedure are highly technical and would take me too far afield if I were to get into them. Rest assured the procedure works, as the SPAUN model evinces, and that modelers are capable of decoding and displaying an "internal," decoded representation of the information encoded within neuron populations in the current iteration of SPAUN (see Figure 2 on page 19).

The second principle determines how transformations are performed on the representations encoded in neuron populations within the model. As stipulated in the NEF, transformations are functions of the variables represented by spiking neuron populations, which are determined by an alternative decoding procedure to that applied in the first principle defining the representations themselves. This is because "when defining a computation we identify transformational decoders that estimate some function, f(x), of the represented quantity. In other words, we find decoders that, rather than extracting the signal represented by a population, extract some transformed version of that signal" (58). The purpose of transformational decoders is to account for the brain's ability to transform (i.e., compute over) incoming information. In other words, unlike representational decoders whose output is a direct, approximate reflection of input information, transformational decoders transform the encoded information into another form, most commonly, into a high-dimensional vector in the case of SPAUN.

Such transformations are integral to cognition, since without transforming incoming information, it becomes useless within the system. For example, if the system is presented with a bus, that information is received in the form of light entering the eye. The information contained in those light rays must then be transformed into a high-dimensional vector representing buses, otherwise it will only be capable of storing raw light ray data, or the activation patterns of photo-sensitive cells in its simulated eye, without understanding that it is looking at a bus. It should be clear from this bus example how the second principle builds on the first: the first stipulates that encoding and decoding procedures are necessary for characterizing neural representations, and the second stipulates that transformations must be performed on encoded information for it to be of any use. It is important to note that these functions can either be determined by designers or learned by the system. One of the virtues of the SPA I have yet to directly address is that it allows for both procedures. Functions analytically determined by modelers can be thought of as characterizing the reflexes and innate knowledge we associate with human nature, whereas learned functions can be thought of characterizing the conditioned and learned responses we associate with nurture. This learning is clearly displayed in SPAUN's performance on the bandit task detailed in the previous section.

To address the philosophical theories of mind underlying the first and second principle of the NEF: they take inspiration from information theory, and computational-representational theories of mind

(CRTMs), specifically symbolicism and connectionism, to characterize the representational and transformational aspects of the mind. It is important to keep in mind that the NEF rejects the metaphors that each of these theories rely on. As far as these metaphors go, symbolicism is a cognitive theory based on a metaphor comparing the mind to computer software (Eliasmith 2003, 1). In other words, the brain computes information based on rules determined by the mind (or software) and written in a "language of thought" (or code) (1). It is important to note that this metaphor is asymmetrical, in that the target analog (the mind) is understood by comparison to a source analog (computer software). Connectionism, on the other hand, relies on a metaphor comparing the mind to the brain, though the neural architectures typically employed by connectionists lack biological plausibility. This is because the nodes, units, and connections at work in connectionist theories and neural networks do not resemble real neurons. In essence, connectionists are truly relying on a metaphor which compares the mind to brain-like behaviour regardless of a model's resemblance to biological brains. This is where the third principle comes in.

Unlike the first two principles of the NEF, the third principle is inspired not by CRTMs, but by dynamicism. It stipulates that dynamical systems theory (or control theory) can be applied to characterize neuronal systems. Symbolicism and connectionism, unlike dynamicism, have played a central role in cognitive theories since the inception of cognitive science (2). This is because the dynamicists' preferred metaphor for characterizing the mind compares it to dynamical systems, one prominent example of which – routinely employed by proponent Tim Van Gelder (Van Gelder *et al.* 1995) – is the Watt governor (3). Such dynamical systems are non-representational, non-computational, low-dimensional, and respond continuously to input information. Thus, dynamicists also argued that any CRTM based on discrete states is prone to mischaracterizations of the mind, since it cannot account for the mind's continuity and timely response to its environment (3). This argument regarding the inadequacies of discrete-state machines, though not entirely ill-founded, is also shared by connectionists (Eliasmith 2001, 4), so it is not truly a concern that distinguishes dynamicism from other cognitive theories. In fact, modern control theory has always played a role in characterizing neurobiological systems (Eliasmith 2003, 12), since they do indeed respond in continuous fashion to their environments, as "moving, eating, and sensing in a constantly changing world are clearly dynamic processes" (Eliasmith 2013, 62). Thus, if a theory is going to capture the mind functionally, and in its entirety, it must take full stock of the dynamical nature of cognitive systems, hence Eliasmith's move to synthesize dynamicism – which was originally touted as a view in contradistinction to CRTMs – together with them. This synthesis requires, however, that we drop the idea that the mind is non-representational, non-computational, and low-dimensional, since these notions are ill-founded, lack sufficient implementation, and contradict core assumptions of symbolicism and connectionism respectively (Eliasmith 2001, 11).

Nevertheless, Eliasmith is adamant that any metaphors underpinning the theories that inspired the principles of the NEF and SPAUN's architecture, are not theoretical arbiters, since none of them alone are able to exhaustively encapsulate the full function of the mind (Eliasmith 2001, 3; Eliasmith 2013, 14). The NEF is then a reframing of a critique and new synthesis of previous cognitive theories by Eliasmith (Eliasmith 2001, 3). For simplicity, I will refer to the Representation and Dynamics Theory (R&D Theory) as a philosophical and normative theory. That is, I will intentionally conflate what Eliasmith distinguishes as R&D Theory and the NEF. The only difference between them, is that the former suggests how neural engineering *ought to be* done, whereas the NEF alone is a descriptive framework determining how it *is* done in the case of models like SPAUN. It is useful to consider R&D Theory for three reasons: it clarifies how the principles of the NEF take inspiration from representational theories of mind and dynamicism; it functions as the philosophical groundwork upon which the NEF is based; and it explicitly incorporates the philosophical perspectives I intend to challenge.

The NEF, via R&D Theory, thus defines the mind as a "complex, physical, dynamic, and information processing system" (Eliasmith 2003, 7). Considering that the brain itself is also taken to be just such a system, this definition equates the mind with the brain, and is therefore a statement of mind-brain token identity theory. This definition also implies monistic materialism (a.k.a. physicalism) since any system in question is ultimately considered to be physical. One might be tempted to think that this theory closely resembles, and is indistinguishable from connectionism, since connectionists rely on a metaphor comparing the mind to the brain. This is not the case, however, so I am now going to clarify how the two differ.

First, as previously mentioned, the NEF does not rely on any metaphor as its theoretical arbiter. This is to say that the identity relation between mind and brain posited by identity theorists is much stronger than the comparative, metaphorical relation posited by connectionists; for identity theorists, the mind is not *like* the brain, it *is* the brain (15). Second, the neural networks typically employed by connectionists rely entirely on learning mechanisms to determine the functions characterizing representational transformations, whereas the NEF allows modelers to determine these functions analytically as well (19). Such analytically determined functions are employed, for example, so that SPAUN knows beforehand that a question mark is to be followed by a response. Third, the NEF differs from connectionism insofar as it takes biology and neuroanatomy much more seriously (18). So, unlike connectionists who are willing to attribute cognition to systems that do what brains do – in a very input-output, behaviouristic kind of way – NEF theorists are only willing to attribute cognition to systems that do what brains do in the same way that brains do them. Thus, NEF theorists are wedded to a specific kind of functionalism[8] which, though it has not yet seen a clear formalization, has nevertheless been dubbed "micro-functionalism" by some of its practitioners. It is important to note that micro-functionalism has seen no clear formalization in the cognitive science literature, rather, it was a term used by Eliasmith at some early point to denote the constrained functionalism that eventually led to the formalization of the NEF. What I offer here is therefore a formalization of micro-functionalism which hitherto has only been an informal label.

Micro-functionalism is born out of Eliasmith's critique of the classical Turing machine style functional equivalence which proposes that it does not provide a definition of computation robust enough

[8] Functionalism: A theory of mind that defines mental states by their causal relations to sensory inputs, behavioural outputs, and other mental states. In other words, functionalists define mental states by their functions: what they do. Multiple realizability – the view that physical systems with differing substrates can realize the same causal relations – is the key component of this theory that leads its proponents to believe non-biological systems can reproduce mental states.

to capture the mind (Eliasmith 2002, 6), since it does not offer a complete description of physical implementations (7). This critique is reminiscent of that made by Ned Block, who proposed that functionalism will invariably lead us to ascribe cognition either too liberally, or too chauvinistically (266). This is because the form of computation performed by Turing machines is defined by the ability to produce an output determined by a machine state table from a given input. By such an unconstrained, behaviouristic definition, however, just about anything can be considered a computer, and by extension, a cognitive system. This could lead to absurd attributions, such as the notion that a gas engine is a computer, and by extension, a cognitive system, because it has determined inputs and outputs. To combat such absurd attributions and restrict the domain of what can be considered cognitive systems, Eliasmith instead proposes:

> […] We need a notion of (Kolmogorov) equivalence based on considerations of computational complexity and computability. Specifying boundaries on temporal and/or computational complexity within which a system must fail to count as cognitive is essential to a useful taxonomy of minds. This is because there is no getting around the fact that in the real world the set of functions realized by a given physical system is going to depend significantly on the physics of that system.
>
> (Eliasmith 2002, 7)

In simpler terms, a gas engine, though complex, does not compare to the sheer complexity of brains which we know minds accompany. Nor could a gas engine reworked to function computationally accomplish cognitive tasks within the same time frame as we do. This is because regardless of the virtues of theoretical Turing-style functional descriptions of physical systems, their physical implementation will always constrain the system's performance (7). One might be able to perform computations using a gas engine, but chances are these computations would take far too long for a functionalist to ever consider it a cognitive system, since it does not meet the demand that the system be able to accomplish everything we can, including producing outputs at a certain rate. This fact about the speed of computation also applies to super intelligent AIs, since they typically perform cognitive tasks such as mathematical calculations much faster than we can, not to mention their lack of general intelligence and perfect memory. This is all to say the concern for complexity and temporal boundaries indeed makes micro-functionalism a far more tenable and much less controversial position than its predecessors, since these restrictions contribute to the overall biological plausibility of any non-biological, potentially cognitive systems it might produce.

Despite micro-functionalism's relatively uncontroversial nature, the main philosophical assumption inherent to this new paradigm is the idea that qualia, and the existence of the self who possesses them, come free with properly structured computation reminiscent of brain activity. While this idea appears at first to be rather banal and uncontroversial, since as far as we know, there are no minds without brains, some controversy lies in the fact that identity theory in tandem with physicalism have thus far "solved" the problem of incongruities arising between the phenomenal experience selves have of freely deciding, and the fact that the brain is a physical system – whose operations are entirely determined by social, environmental, biological, and quantum factors – by regarding the phenomenal experience as a falsehood, or a "user illusion," and by taking a compatibilist stance which reduces freedom to self-controlling or learning mechanisms which machines also possess. In other words, because the mind is believed to be identical with the brain, the brain is a physical system, and the physical is either deterministic or indeterministic, there appears to be no purpose left for us: brains will do what they are determined to do or they will act randomly, and any sense we have that we are making decisions freely is an illusion attributed to our own ignorance resulting from the sheer complexity and unpredictability of the brains we all possess, hence why I have spent the entirety of this chapter characterizing SPAUN. This new paradigm would have us believe that SPAUN could one day (upon its completion) possess qualia and free will because they come free with properly structured computation that can be realized in brain activity, though they would be considered illusory. After all, SPAUN is, by any functional account, a brain.

If you are a functionalist, identity theorist, or physicalist, you are probably wondering why this is even a problem. Besides the fact that determinism nor indeterminism do not take full stock of our phenomenal experiences, compatibilism only offers a "modest" free will which requires a restructuring of our views on autonomy, moral responsibility, blame, and justice. Typically speaking, we tend to believe that criminal actions are deserving of punishment because the criminal in question is a free agent who could have done otherwise if they so wished. Nevertheless, we also tend to believe that though they could have done otherwise, external factors like upbringing or mental disability may also influence criminal behaviour. Depending on how restricting we believe their circumstances to be, we may judge that their influence is impactful enough to be exculpatory, and that the criminal in question is rather in need of rehabilitation or medical intervention, as in cases of criminal insanity, for instance. Our typical notions of free will and moral responsibility, then, are decidedly incompatibilistic, in that a criminal act is either free and therefore punishable or determined and therefore not punishable. Nevertheless, the new paradigm consisting of physicalism and identity theory would have us believe that all criminal acts are determined, since all criminals have brains, and because brains are physical systems, they are determined to produce criminal behaviours irrespective of whether a criminal is of sound mind or not. In other words, if everyone has a brain and no one with a brain could have done otherwise, then evil individuals do not choose evil freely, making the idea that they deserve blame and punishment for their evil deeds obsolete.

No thinker better represents this new paradigm than Daniel Dennett. His compatibilistic attempts to reconcile free will with determinism are compelling enough philosophically to provide exactly the kind of support AI enthusiasts require to justify their belief that SPAUN, for instance, has, or one day will have a mind. Furthermore, unlike some of his contemporaries (e.g., Sam Harris), Dennett is adamant that compatibilism requires no restructuring of our current criminal justice systems. Because of the prevalence of his philosophy among cognitive scientists, the next chapter is devoted to characterizing and critiquing his views. The purpose of this chapter is to show that despite Dennett's insistence that free will and determinism are compatible, and that compatibilism requires no restructuring of our concept of moral responsibility, this concept is in fact inextricably linked to classical incompatibilist[9] views on free will. That said, Dennett's philosophy goes hand in hand with theoretical neuroscience, so I also analyze discoveries

[9] Incompatibilism: The view that determinism is not compatible with free will.

in neuroscience which at first glance appear to challenge the notion that we have classical, or “ambitious” free will. The purpose of these sections is to show that the results of these studies are ultimately inconclusive, thereby establishing the need for a philosophical approach to this subject.

Chapter 2

A Critique of Dennett's Compatibilism

It is important to note that Dennett's views are highly nuanced, his writing is prolific, and his abundant use of intuition pumps can make his position difficult to pinpoint at times. Most of my critiques will be focused on the arguments he has laid out in *Elbow Room* and *Consciousness Explained*, with a few exceptions. I also contrast his view with that of Sam Harris, since Harris is one of Dennett's contemporaries, and a neuroscientist with much to say about determinism and moral responsibility. His position is also much simpler and easier to interpret than Dennett's. Moreover, Harris brings up multiple points in his discussions with Dennett that parallel some of my own critiques. Though we might share concerns regarding Dennett's position, Harris, like Dennett, is also a determinist/materialist, so our positions ultimately diverge. I bring Harris into this discussion to highlight and clarify where I believe Dennett leads the discussion astray, in a way that diverges from mine and Harris' viewpoints.

The publication of *Elbow Room* in 1984 marks Dennett's first attempt at expounding his view that free will is compatible with determinism, so long as we conceive of free will in a particular way. I have already alluded to what conception of free will Dennett champions in the previous chapter, and I will be further characterizing it in this one, but for the moment an explication of his motivation for this work is in order, since it will help me illustrate where his view reaches a dead end, at least explanatorily. His stated motivation for developing his view is that philosophy is prone to make arguments for "magical" and untenable classical notions of free will out of fear of bugbears: non-existent entities and analogies which are concocted to coerce us through fear into believing that an alternative is abhorrent (Dennett 1984, 4). The bugbears, or "bogeymen" he lists are:

- The Invisible Jailer: determinism likened to being in an invisible prison run by an invisible jailer (7).

- The Nefarious Neurosurgeon: determinism likened to having your actions determined by a neurosurgeon who has inserted electrodes into your brain. Versions of this bugbear include the Hideous Hypnotist, and the Peremptory Puppeteer (8).
- The Cosmic Child Whose Dolls We Are: determinism likened to being a toy under the control of a childish, capricious, God-like entity (9).
- The Malevolent Mind Reader: determinism likened to being stuck playing a game where the opponent can predict your every move (10).

It is important to note that Dennett contributes these bugbears himself, and they are not explicitly mentioned by classical philosophers, though he believes many of their arguments in favour of free will rely on allusions to them. Dennett is right to propose that because these entities are more-than-likely non-existent, we should not believe in free will simply out of fear that the alternative of being determined, biological machines would be akin to being locked up by the invisible jailer, being probed by the nefarious neurosurgeon, being the plaything of the cosmic child, or having our actions predicted by the malevolent mind reader. After all, from the scientific evolutionary perspective of naturalism, there is no intention behind the physical forces that determine biological activity. Naturally, he moves on from these bugbears to more tenable concerns regarding sphexishness: the condition of being a biological mechanism or automaton; much like the sphex wasp, whose actions are completely determined by biological imperatives. He argues that two different unsettling notions present themselves when considering the idea that we are sphexish:

- The Disappearing Self: that determinism undermines the purpose of the self. The sphex can function in the world and continue its survival without consciously understanding anything or having a sense of self, so why not us too (13)?
- The Dread Secret: that proof of determinism would be like opening Pandora’s box, and all deliberation would cease, since the outcome is already determined, and all our debates about the way things ought to be amount to nothing (14).

Compared to the bugbears previously mentioned, these concerns are far more reasonable. In fact, a good chunk of *Elbow Room*, and even subsequent works (e.g., *Freedom Evolves*), are devoted to abating them. At this point, I want to draw your attention to the problem of the disappearing self, as I believe it constitutes the very point at which Dennett's arguments reach a dead end. By that I mean it is a problem Dennett never provides an adequate solution for, though he attempts many over the course of his publications, including *Elbow Room*, *Consciousness Explained*, and *Freedom Evolves*. This problem also surfaces in multiple forms: from the philosophically motivated P-zombie[10]argument made by David Chalmers, to the discovery of blindsight by neuropsychologist Lawrence Weizkrantz. Before I get into those details, however, I will first provide a brief exposition of how he deals with the dread secret problem, since it also constitutes a reasonable concern, and unlike his solution to the problem of the disappearing self, Dennett manages to provide an adequate solution for it.

Dennett solves this problem rather simply: deliberation, though determined, still affects our actions, and by extension, physical outcomes (104). In other words, it is worth deliberating whether that deliberation is done freely or not. To say any deliberation is rendered futile by determinism is to say that any attempt at calculating 2 + 2 is futile, since it is already determined whether I produce the correct answer or not: I might as well guess. One can plainly see the error in this line of thought. One needs to do the calculation to get the correct answer, or otherwise get incredibly lucky. Thus, even if it is determined that one should do the calculation and get the correct answer, that deliberation was not futile, since it still contributes to the outcome of this scenario. Implied in his solution to this problem is the idea that indeterminism[11] does not constitute an adequate alternative to determinism, since it does not preserve a form of free will worth wanting (2). Indeed, a world where our actions are not determined by any laws, but by randomness alone, is not a world in which we are free either. Within such a world, deliberation would in fact be a futile

[10] P-zombies: Otherwise known philosophical zombies. P-zombies are physical twins that possess no qualia. They were introduced by David Chalmers in 1997 as a challenge to materialism, taking after Frank Jackson's knowledge argument that qualia are non-physical and cannot be explained by physical accounts.

[11] Indeterminism: The idea that not all events are wholly determined by antecedent events. Heisenberg's Uncertainty Principle stipulates that quantum mechanics are indeterministic, and thus probabilistic. Nevertheless, determinism appears to reign at larger scales of observation, such as our everyday experience.

endeavour, since no amount of deliberating could ever affect the random outcome of an action. The deterministic evolutionary perspective of naturalism is well-suited to explaining how deliberation might come about by natural selection, since deliberating systems possess an adaptive edge which contributes to their survival (94, 98). Thus, I take no issue with the "just so story"[12] explaining the evolution of deliberation he offers as a solution to this particular problem.

So much for the dread secret problem, but what about the problem of the disappearing self? In *Elbow Room*, Dennett provides a naturalistic account for the existence of selves and their conscious experiences of making decisions (76). He also revamps this same argument in *Freedom Evolves* nearly 20 years later. For simplicity and brevity's sake, I will mainly be focusing on the account provided in *Elbow Room*, though I will also provide a brief explanation of how the critique I offer applies to his later account. That said, the first order of business is that of characterizing the self as it is defined by classical thinkers, since I will be defending a similar one, and this is a definition Dennett takes issue with. To start with, he provides a less philosophical account of classical views on free will by considering the notion that we could have done otherwise (131). The idea, roughly stated, is that given the exact same physical state as when a decision was made, we could have made an alternative decision (132). This "could have done otherwise," or "can do otherwise" indeed reflects the "ambitious" variety of free will classical philosophers tend to promote. Their conception of the self, then, is that of an entity not entirely restricted in their actions by deterministic physical laws, nor existing in an indeterministic world where their seemingly free actions can be attributed to random events. Dennett then correctly identifies the conception of free will offered by classical philosophers as agent-causation, and believes the definition provided by Chrisholm to be its clearest expression: "If we are responsible… then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain events

[12] "Just so story": A preliminary and untestable narrative account of cultural practices, biological trait, or behaviour. These are often used as explanations in cases where empirical evidence is either non-existent or inconclusive.

to happen, and nothing – or no one – causes us to cause those events to happen" (76, quoted from *Human Freedom and the Self*, Chrisholm 1982, 32).

The concept of unmoved movers harkens back to Aristotle who, in *Physics* and *Metaphysics*, describes the prime mover or originator of all motion, to be an unmoved mover: an infinite motive force without magnitude (*Physics*, Aristotle 1941, 380, 390). Aristotle also describes the prime mover as a final teleological cause, and as God (*Metaphysics*, Aristotle 1941, 879-880). This concept later resurfaces in the philosophy of Thomas Aquinas, a 13th Century medieval philosopher/theologian, who, in *Summa Theologica*, presents five famous arguments for the existence of God, the first three of which take inspiration from Aristotle's prime mover concept (Aquinas 1947, 15). Roughly speaking, both Aristotle and Aquinas posit an uncaused God to avoid positing an infinite causal regress which they both believe to be absurd (*Physics*, Aristotle 1941, 367; Aquinas 1947, 15). This is all to explain Chrisholm's statements regarding responsibility. According to him, unless our actions are uncaused in a manner akin to God, then they are not creative; therefore we cannot be held responsible for them.

Unfortunately, while I remain agnostic about the existence of God, and do not deny the possibility of its existence, these tales about prime movers, God, and souls rely on faithful hypotheses that do not mesh well with the naturalism inherent to the scientific method. Cognitive science, being that it is a science, cannot rely on these arguments, rather, it needs to construct its own naturalistic account. On the notion that naturalism provides the proper arena for debates around freewill, I agree with Dennett. This is one reason I have chosen to promote "ambitious" free will instead of the "magical" free will characterized by some Ancient and religious classical accounts. In other words, though both "ambitious" and "magical" free will are characterized by unmoved movers, "ambitious" free will can be maintained without recourse to God, or souls, since it essentially posits the self as the end of the causal line. This brings me to the other reason I do not promote "magical" free will: the existence of both souls and God detracts from the notion that the self-determining of selves is uncaused. After all, if self-determining selves are caused by souls or God, their genuine status as self-determining can be called into question on the grounds that the self-determining process is itself determined.

Dennett critiques Chrisholm's account, before moving on to produce his own, both of which I will summarize and discuss. In short, Dennett identifies "the intuitions that support Chrisholm's vision of the self as unmoved mover to be a sort of cognitive illusion" (Dennett 1984, 76). He attributes this illusion to the fact that, much like in the case of Aristotle and Aquinas, who posit an uncaused cause to circumvent an infinite causal regress, we posit ourselves as uncaused causes of our own actions because of the inscrutability of the causal chain that truly produces our behaviour (77). According to Dennett, this inscrutability, weighed against the experiences we have of coming to decisions, results in "the hypothesis that there are no causes" (77); a.k.a. agent-causation. In other words, we fill the gaps in our knowledge about the physical, causal mechanisms that, according to Dennett, truly determine our behaviour with the "magical" unmoved mover out of fear that we are not in our own control, and by extension that we cannot justify our belief that we and others are morally responsible (78). He further argues that these convictions are based solely on the false hypothesis that no naturalistic account can be provided for the existence of selves (80).

At this point, he moves on to produce just such an account. To start, he identifies the self as the locus of self-control (81). The notion of control is key to understanding Dennett's characterization of selfhood, since it contrasts with the notion of creation which is key to understanding its classical characterization. According to Dennett, the notion of control is all we need to safeguard our sense of free will and moral responsibility (82). This notion also comes with the added bonus that it does not require any suppositions related to magical, unnatural, or non-physical entities like unmoved movers. While he does not provide the same kind of detailed "just so story" in this case as he does when dealing with the dread secret problem, or at least not in *Elbow Room*, his argument is still naturalistic. In essence, he posits that a self-control mechanism in the brain evolved over time by natural selection, and that the purpose of this self-control mechanism is, among other things, to cut short the deliberative process, since without such mechanisms, deliberations would never end, and nothing would ever get done, making the ability to deliberate not so adaptive after all (164). That is, in order for us to deliberate the appropriate amount, a meta-level self-monitoring system like consciousness needed to develop on top of the ability to deliberate so as to keep it in check.

The "just so story" he offers in *Consciousness Explained* to explain the evolution of consciousness picks up and builds upon a hypothesis developed by psychologist Odmar Neumann, who proposed that "orienting responses" developed at some point in evolutionary history so animals could deal with dangerous predators (Dennett 1991, 180). These "orienting responses" are juxtaposed with the "autopilot" we are all quite familiar with (180). Dennett hypothesizes that such orienting responses functioned as an evolutionary precursor for consciousness, and that over time more and more species developed more and more orienting responses because they constitute an adaptable trait (180). The result is that we as a species have a system in place to anticipate potential, future dangers, and according to Dennett, this is consciousness (181). Another version of this story is provided in *Freedom Evolves*, though this version relies on a metaphor comparing the orienting responses of mammals like us to the avoidant behaviour of "avoiders" which develop through an analogous evolutionary process in Conway's Game of Life (Dennett 2004, 51).

The Game of Life is software constituting a Democritean toy model consisting of a two-dimensional grid of pixels which can either be ON or OFF. The "physics" of this toy model are accounted for by programming simple rules into the system. Each pixel has eight neighbouring pixels. If exactly two pixels neighbouring a pixel in question are in the ON state, the state of the pixel in question remains unchanged, whether it is currently ON or OFF. If exactly three pixels neighbouring a pixel in question are in the ON state, the pixel in question turns ON. The pixel in question will turn OFF in all other circumstances (36). Given these simple rules, pixels in the Game of Life come to represent cells in a biological system. Over time, pixels come together and form constructs of various types, each of which display unique behaviours. Examples include five-pixel gliders that "glide" about the two-dimensional plane, and seven-pixel eaters that "eat" any other construct they happen to encounter (39). Using the Game of Life as an analogous example, Dennett proposes that at some point avoidant behaviours necessarily developed during the evolutionary process, since otherwise all constructs besides eaters would have died out at some point, leaving only the eaters to eventually die out as well (43). I intend to provide a critique of these “just so stories,” but for now I will turn to Dennett’s critique of the classical, “ambitious” conception of free will. The reason for this is that my disagreement with Dennett rests firmly on his conviction that we have free will despite his insistence that we could not have done otherwise. Thus, I need to lay out his justifications for this so I can demonstrate the tension between these convictions.

After providing his naturalistic accounts, he goes on to critique the classical reliance on the notion that we are morally responsible because we could have done otherwise (Dennett 1984, 131). In this critique, he proposes that classical philosophers have misinterpreted what people mean when they say, “I could have done otherwise.” As previously stated, philosophers have typically interpreted the meaning of this phrase as “if we could rewind time so that I could be placed in the exact same scenario as before, I could make a different decision than I did.” Dennett’s interpretation of the meaning of this phrase is rather that one is inquiring as to the potential causes of an incident so that they can learn from it and improve their actions in the future (142). To say we could have done otherwise is then not to make some metaphysical assumption about our existence as unmoved movers, it is only to say we realize there was another way to act in those circumstances, and that if put in a similar scenario in the future, we would not act the same way. Morally speaking, the result of this line of reasoning is that taking responsibility for our actions amounts to realizing that we can learn from their consequences and improve ourselves in the process (143). Thus, what exculpates the criminally insane in this framework is that they lack one of the two essential features of control: they either lack the brain mechanism responsible for deliberation, or they lack the brain mechanism responsible for self-control or learning. What about the average criminal though? I previously explained how applying this logic in the case of the average criminal entails the restructuring of our views on justice. This is one point I do not believe Dennett provides an adequate retort for, though he does address it, so I will quickly cover his response to it.

Dennett considers first and foremost the most explicit response to criminal conduct by an agent we take to be morally responsible: punishment (158). Dennett believes that penal institutions remain justified even though the average criminal – who we take to be a self-controlling deliberator like us – could not have done otherwise. He justifies this by stating that we do not live in an ideal world where everyone respects the law, and by citing Paul Gomberg's notion that legal punishments for the average criminal are justified insofar as they "may cause him to shed an undesirable trait, and this is useful regardless of whether this trait is of his making" (164, quoted from Gomberg 1978, 208). This is exactly the kind of response I find Dennett's philosophy renders obsolete, the challenge being that compatibilism proposes "one acts freely and responsibly just so long as one does what one decides, based on what one believes and desires" (83), but one's desires and decisions are determined by their brain, whose operations themselves are either exhaustively determined by biology and the environment, or they are random occurrences. How then, can we take credit, morally speaking, for acting in accordance with just desires, or making just decisions, when it seems we had no hand in constructing them in the first place? We are now considering the converse scenario of course: one in which we blame people for acting in accordance with unjust desires and making unjust decisions.

This response is inadequate because it overemphasizes the idea that we punish criminals for the purpose of conditioning, since we are also prone to seek retributive justice. That we often do not hold any sympathy for criminals and seek retributive justice as payment for their misdeeds suggests adopting compatibilism requires that we reconsider our concept of justice to account for the rehabilitative needs of criminals regardless of their status as responsible agents. You may be wondering about the distinction between what is and what ought to be at this point. Just because we do seek retributive justice, does that mean we are right to do so? I argue that we are at least sometimes correct in this, but that of course depends on the notion that at least some evil doers choose to do evil deeds freely, and freely shirk any opportunities to improve their characters. If the door is open for dualist interpretations of Being that justify beliefs in "ambitious" free will, which I argue it is, then the intuitions driving our desire for retributive justice are often justified, and the criminal justice system, as it currently stands and despite its many flaws, is acceptable in this regard.

Now, the argument that compatibilism entails accounting for the rehabilitative needs of criminals regardless of their status as responsible agents reflects the challenge Harris offers with respect to Dennett's belief that the criminal justice system requires no reform under his compatibilist framework. Unlike Dennett, whose compatibilism is grounded in a soft determinism, Harris's views are grounded instead in hard determinism, meaning he believes determinism is incompatible with free will. Thus, his challenge parallels my own and that typically posed by hard determinists, which is best illustrated by Paul Edwards:

> Let us suppose both A and B are compulsive and suffer intensely from their neuroses. Let us assume that there is a therapy that could help them, which could materially change their character structure, but that it takes a great deal of energy and courage to undertake the treatment. Let us suppose that A has the necessary energy and courage and B lacks it. A undergoes the therapy and changes in the desired way. B just gets more and more compulsive and more and more miserable. Now it is true that A helped form his own later character. But his starting point, his desire to change, his energy and courage, were already there. They may or may not have been the result of previous efforts on his own part. But there must have been a first effort, and that effort at that time was the result of factors that were not of his making.
>
> (Dennett 1984, 83 quoted from Edwards 1961, 121)

Applying this line of reasoning to the case at hand, if one were not to shed their undesirable traits after suffering the punishment of incarceration, then incarceration does neither the criminal, nor society, any good, besides satisfying our supposedly primitive desire for retribution. It would rather seem that because the criminal in question was determined not to learn their lesson by our current means of teaching it, they are not at fault, rather, our current means of teaching the lesson are faulty. This consideration leads hard determinists like Harris to propose that our notion of retributive justice is obsolete, since it does not seem conducive to the conditioning required for rehabilitation, as Dennett suggests. Dennett does address concerns about hard determinism, though his rebuttal is less than adequate. He argues that responsible selfhood is acquired, and that once acquired, though we might not have complete control over our actions, we do have some (85). Those with the ability to self-evaluate will find themselves with more control over their actions and generally speaking, the better the self-evaluator the better their character (87). The obvious problem with this response is that it begs the question. The challenge posed by Edwards, Harris, and myself is that in a deterministic framework, the degree to which we self-evaluate is also determined by every link in the physical causal chain preceding ourselves. If a self-evaluator finds they do not have it in them to act justly, then this was not of their own doing.

Why does it matter whether we punish or rehabilitate criminals? What relevance does this have to the topic of whether SPAUN is conscious or not? Well, from all I have said, it should be clear that we often do not incarcerate criminals because we think they will come out the other end of prison reformed. We do so because we seek retributive justice. We believe they deserve to pay for what they did, and the only reason we believe this is because we recognize that they could have done otherwise, in exactly the way classical philosophers have historically interpreted this phrase. I suspect that some physicalists would readily concede that if a hypothetically completed and physically realized SPAUN-like automaton were to commit a crime it would deserve the same treatment we do, and should therefore be punished for its transgression; however, I find it most likely that this intuition would be challenged if the scenario were to actually play out. This is because, unlike our intuitions about human beings and their ability to have done otherwise, we believe machines behave mechanistically, that they have no ability to have done otherwise, and therefore that they deserve no punishment for their misdeeds. After all, if a SPAUN-like automaton were to commit a crime, our response would probably not be to throw it in prison, but either to destroy it or recall it for maintenance. Given the critiques I have offered of Dennett's position that adopting a deterministic and compatibilistic framework in which we could not have done otherwise does not require a restructuring of our views on moral responsibility and justice, it should be clear that in fact these views are intimately related to the notion that we could have done otherwise.

You may have already caught on to how the critique advanced by hard determinists relating to moral responsibility might also serve as a critique of Dennett's "just so story" explaining the evolution of consciousness. If we are to liken consciousness to a self-monitoring, enhanced orienting response that enables our avoidant behaviours, as Dennett suggests it is, then we cannot do otherwise than avoid. Avoidant behaviour, if determined and not freely chosen, is then not something for which we can take credit, or lay blame in the case of its absence. Moreover, the self, defined by Dennett as the locus of self-control, ascribes too little causal properties to agents to justify our attitudes regarding retributive justice. The criticism I am advancing, then, is essentially that, given the supposed validity of intuitions guiding our judicial and penal institutions, the self must be what could have done otherwise in the classical sense. Thus, I am taking issue not with compatibilism or determinism per se, but with hard determinism and Dennett's particular brand of soft determinism, since it is still too "hard" to justify intuitions regarding retributive justice.

The Self Disappeared

Besides the moral side to this argument, Dennett's soft determinism is also too "hard" to take full stock of our experiences of freely deciding, leaving an explanatory gap. He suggests instead that this phenomenal experience is merely a cognitive illusion. I am calling into question the effectiveness of Dennett's naturalistic account as a solution to the problem of the disappearing self, because in order to be a responsible self, one must be able to freely decide in exactly the way they experience themselves deciding. This results in the aforementioned tension that arises in Dennett's philosophy: in the one hand he holds our institutions of justice, and in the other he holds a compatibilistic view of the self and free will that cannot ground them. He cannot have it both ways. I surmise, as neuroscientist Benjamin Libet did, that "Great care should be taken not to believe allegedly scientific conclusions about our nature that depend on hidden ad hoc assumptions. A theory that simply interprets the phenomenon of free will as illusory and denies the validity of this phenomenal fact is less attractive than a theory that accepts or accommodates the phenomenal fact" (Libet 2004, 155). I admit this statement is somewhat hypocritical coming from Libet, for reasons I will clarify in later sections where I analyze his experimental data. Regardless, whether our phenomenal experiences of freely deciding are illusory or not, Dennett's naturalistic account does a rather poor job of explaining its *purpose*, and so his account virtually guarantees that the self disappears.

What I mean by this is that Dennett never takes seriously the possibility that evolution could have produced a deliberating mechanism, and a mechanism to monitor the deliberating process, without any "illusion" or qualia accompanying them at all. That is, his naturalistic account explains the evolution of deliberative mechanisms and control mechanisms, not *deliberators* and *controllers*. This is because he is operating within a physicalist framework where selfhood and phenomenal experiences come free with the computational organization that gives rise to these mechanisms, and where any first-person phenomenal experience we might have, such as the experience of freely deciding, is merely an illusion if it does not correlate with some operations within the physical system of the brain. The question which goes unanswered is, of course, "why?" This brings me to the discovery of blindsight by Weizkrantz in 1970 (Ramachandran 2011, 61), which I will now briefly recount since its discovery led neuroscientist Vilayanur Ramachandran to ponder this very question, and it will help me pinpoint the explanatory gap in Dennett's naturalistic account.

When assessing the condition of a patient who had suffered substantial damage to the visual cortex, known only as GY, it was found that he had become completely blind in his right visual field. In the course of testing his vision, Weizkrantz shone a light on the right side of GY's visual field and asked him to reach out and touch it. GY of course protested that he could not see on that side, but despite his protest, Weizkrantz asked him to try anyway. Surprisingly, GY was able to reach out and correctly guess the location of and pinpoint the light spot without the ability to consciously perceive it. GY was adamant that he was guessing, but repeated trials showed that this had been no fluke on his part. He was able to correctly pinpoint the location of the light spot every single time (61). Ramachandran's explanation for this seemingly miraculous phenomenon is that while the “new pathway” of the visual cortex was damaged, resulting in a lack of conscious visual processing, the “old pathway” through the superior colliculus to the parietal lobes was undamaged, and still processing the visual information unconsciously (61). According to him, this indicates consciousness accompanies only the “new pathway” of the cortex (61). In his address at the Reith Lectures in 2003, he expands on the implications of this discovery:

> It's as if even though GY the person, the human being, is oblivious to what's going on, there's another unconscious zombie trapped in him who can guide the hand movement with uncanny accuracy. This explanation suggests that only the new pathway is conscious - events in the old pathway, going through the colliculus and guiding the hand movement can occur without you the person being conscious of it! Why? Why should one pathway alone or its computational style perhaps lead to conscious awareness, whereas neurons in a parallel part of the brain, the old pathway, can carry out even complex computations without being conscious. Why should any brain event be associated with conscious awareness given the "existence proof" that the old pathway through the colliculus can do its job perfectly well without being conscious? Why can't the rest of the brain do without consciousness? Why can't it all be blindsight in other words?
>
> (Ramachandran 2003)

Why indeed. The challenge this inquiry poses to Dennett’s framework parallels concerns related to the hard problem of consciousness. There is a subtle difference between them, however: where the hard problem does not constitute an objection to Dennett’s account and methodology, since as far as we know, this problem is empirically insurmountable, and expecting Dennett to provide an account of this or a method of solving it would be quite unreasonable; Ramachandran calls for an explanation of what evolutionary advantage consciousness awareness might impart. Would not the same brain mechanisms be just as adaptable were they not accompanied by conscious, qualitative experiences, not to mention the supposed “user illusion?” This line of reasoning will inevitably lead me to consider Chalmer’s P-zombie problem, and I will shortly, but before doing so, it is important that I provide a brief overview of Dennett’s proposed method of cognitive investigation and his theory of consciousness, since the P-zombie problem functions as a challenge to his method as well as his naturalistic account.

In fact, he has gone through great pains to construct a third-person method of cognitive investigation called *heterophenomenology* – which treats the phenomenal experiences recounted by subjects as fictions to be correlated with neuroscience data (Dennett 1991, 81) – as well as a third-person, empirical theory of consciousness that opposes Cartesian Materialism called the Multiple Drafts Model (MDM) (111). The MDM proposes that "Feature detections or discriminations only have to be made once" (113); which is to say that different brain areas are responsible for interpreting different detected features, that do not then need to be reinterpreted by a "master" discriminator. Further implications of this model are that the onset of activity in one brain region does not guarantee simultaneous conscious awareness of its correlated content, that it is incoherent to ask when we become conscious of detected features, and that the supposed narrative structure of conscious experience is an illusion resulting from the multiplicity of competing and continually edited interpretations of competing brain areas (113).

I admit that the MDM is a much stronger theory than its classical alternative, since, unlike its counterpart, it has a basis in modern neuroscience. Although it may serve as a much more sound and explanatorily powerful theory than its predecessor, it does not prove all that Dennett claims it does. He claims, for instance, that Cartesian Dualism is dead, but he only succeeds in providing an alternative to Cartesian Materialism, and since Cartesian Materialism and Cartesian Dualism are two separate ideas, disproving one does not necessarily disprove the other. In other words, though he is correct that there is no "headquarters" for consciousness in the brain, our first-person conscious experience is still Cartesian Theatre-esque. Thus, the MDM does not effectively guarantee Cartesian Dualism's falsehood. The justification for the claim that Cartesian Dualism is false rests not on the validity of the MDM and the invalidity of Cartesian Materialism, but on the notion that the Cartesian Theatre is illusory. Why, then, does the "illusion" exist in the first place? This is the explanatory gap I was referring to earlier. Calling the Cartesian Theatre-esque nature of our first-person experience an illusion by adopting a third-person perspective does not explain the first-person perspective. If anything, it explains it away. Interest in the problem of the disappearing self is therefore not solely motivated, as Dennett suggests, by a fear that determinism leaves the self without purpose (though it does), but also by the reasonable concern that it leaves the existence of the conscious experience of the self without adequate *explanation*.

This brings me back around to concerns regarding Dennett's heterophenomenology: his proposed method for phenomenological inquiry. As previously mentioned, this method treats the phenomenal accounts of subjects as though they are fictions: "[Heterophenomenology] involves extracting and purifying texts from (apparently) speaking subjects, and using those texts to generate a theorist's fiction, the subject's heterophenomenological world. This fictional world is populated with all the images, events, sounds, smells, hunches, presentiments, and feelings that the subject (apparently) sincerely believes exist in his or her (or its) stream of consciousness" (98). Given that qualia and phenomenal experiences are subjective and lie outside of the third-person purview of empirical investigation, this concession seems like a necessity. Unfortunately, it devalues phenomenal experiences, since from the first-person perspective, they are not fictional entities, but the most real entities in existence. According to Dennett, this concession secures a method of inquiry which is "neutral with regard to the debates about subjective versus objective approaches to phenomenology, and about the physical or nonphysical reality of phenomenological items" (95).

He also claims heterophenomenology does not purport to solve the P-zombie problem introduced by David Chalmers nor dismiss it (95). The latter claim is true, but the former is not, and this has everything to do with how hard Dennett's brand of soft determinism happens to be. For instance, he argues that heterophenomenology, though not providing a solution, circumvents the P-zombie problem since from the heterophenomenological perspective, given that P-zombies are behaviourally equivalent to actual agents, "there is surely nothing wrong, nothing nonneutral, in granting zombies a heterophenomenological world, since it grants so little" (95). The problem here is that there is clearly something wrong in granting them this, regardless of how little is granted: they do not actually possess what heterophenomenology grants them, i.e., phenomenal experiences (even fictional ones). That heterophenomenology would have us potentially *concoct* fictions to be analyzed I take to be a sign that this method is deeply epistemically flawed.

I want to make it clear that though I believe P-zombies constitute a real possibility, I do not believe they are real. This is because, given our own first-hand conscious experiences, and the knowledge that we are biological beings, it is not a far stretch to assume that other biological beings also possess conscious experiences. This is to say that though I am challenging physicalism, I am not, by extension, challenging identity theory. I bring up these issues only to reframe them in a discussion about SPAUN, or a hypothetically completed and physically instantiated SPAUN-like entity. Such an automaton could indeed constitute the realest and most proximal example of a P-zombie, though it of course would not be physically equivalent to us, only functionally so (assuming it is composed of silicon[13], rather than carbon). That is, I am truly considering here how heterophenomenology deals with what could turn out to be *functional* zombies, or F-zombies, if you will. To clarify, the only difference between P-zombies and F-zombies is the physical substrate of the system in question, so whereas P-zombies constitute a physical twin devoid of qualia, F-zombies do not constitute physical twins, since the physical substrate of F-zombies differ from biological beings.

On top of potentially granting F-zombies fictional phenomena they do not possess, the claim that heterophenomenology is neutral regarding the physical or nonphysical reality of phenomenological items, though true, is not necessarily desirable. For instance, when considering the nature of intentional objects, he states that they are nothing and are made of nothing, just like fictional objects; they are abstract, not concrete (95). This of course implies that they are nonphysical and unreal. Such a conviction shows that while heterophenomenology potentially grants too much to F-zombies, it certainly grants far too little to biological subjects. For the subjects themselves, there is nothing more real than their first-person conscious experience. These problems with heterophenomenology stem from the assumption of physicalism and its

[13] Silicon: The chemical element with atomic number 14. It is the most realistic option for constituting non-biological entities with conscious states. This is because, just like carbon molecules, silicon molecules have 4 unpaired electrons in their outer orbital resulting in them having similar binding properties.

conclusion that the first-person, “abstract,” conscious experience of subjects is somehow subordinate to, or “less real” than the third-person, “concrete,” observations of scientists who observe their behaviour, hence enabling the heterophenomenologist to discount the authority subjects have regarding their own experiences. Ironically enough, the criticism I am currently making in many ways reflects Dennett’s own criticism of dualists: that they have given up on investigating the mind (37). While Dennett’s heterophenomenology does offer us a means of investigating the mind, and so clearly he has not given up on that front, the assumptions underlying his method make it incapable of investigating the mind as it is truly experienced, namely, as reality, and therefore he has given up on investigating the mind as it truly is.

To briefly recapitulate, thus far, in the course of this chapter, I have laid out Dennett’s naturalistic account of the existence of selves, his compatibilist account of free will, his empirical theory of consciousness (MDM), his empirical methodology (heterophenomenology), and his criticisms of dualism. I have offered critiques of each of these and in doing so have demonstrated where they faulter and that the assumptions underlying them are worth challenging. Dennett, however, is not the only one with a naturalistic perspective on such things. I mentioned earlier that Harris provides a much more streamlined perspective rooted in hard determinism. It is important, then, that I consider his arguments as well.

A Critique of Harris’ Determinism

As previously mentioned, Harris poses the same challenge to Dennett’s compatibilism as I and Edwards do. When raised, the challenge reveals a tension between Dennett’s views on punishment, and his conviction that we could not have done otherwise. No such tension presents itself in Harris’ framework, since it is decidedly deterministic and incompatibilistic. That is, unlike Dennett, Harris believes that because the sense of self and free will are both illusory, the punishment of incarceration within the criminal justice system only makes sense so long as it functions as a deterrent for criminal behaviour, and so long as it serves the betterment of society at large (Harris 2012, 40). Nevertheless, that the criminal justice system should dole out such punishments in retributive fashion he finds obsolete (40) because of the supposed fact that the self is illusory, that it could not have done otherwise, and that neuroscience data supposedly proves this (18). That the threat of imprisonment acts as a deterrent for criminal behaviour for most people I find to be a rather uncontroversial point, but any exceptions to this uncontroversial rule are left unaccounted for. Harris concludes that our criminal justice system needs reform, and argues the ultimate goal of the system should be rehabilitation, rather than punishment (40).

This is all grounded on a hard determinism which Harris justifies by citing neuroscientific evidence, specifically the Libet experiment performed in 1985 (16). This is not his only justification for his belief that free will and the self are illusions, however, as unlike Dennett, Harris believes the illusion of the self can be overcome through meditative practice and the achievement of nondual awareness (Harris 2014, 126). In fact, he argues "The illusion of free will is itself an illusion" (Harris 2012, 43) resulting from a misinterpretation of our own phenomenal experiences. A full-fledged critique of his interpretation of the effects and aims of meditative practices is far beyond the scope of this thesis. I will briefly comment, however, that Eastern spiritual traditions and religions are fragmented. Depending on which proponents of nondualism one looks to, it is debatable whether the self-transcendence promoted by Harris can be considered the end goal of meditative or spiritual practice, and furthermore whether self-transcendence truly amounts to shedding the "illusion" of free will. For example, the Kyoto School[14] philosopher Nishitani Keiji, who took inspiration from the teachings of Dogen – an ancient Buddhist teacher of the Mahayana tradition introduced by Nagarjuna – likens nirvana or spiritual enlightenment with the achievement of "Buddha-Nature," (Nishitani 1982, 264) which is described by Dogen as a form of self-transcendence resulting in a loss of ego but also securing absolute freedom for the enlightened:

> At another time, as Nagarjuna was sitting, he manifested a body of absolute freedom – it was just like the round full moon. Not a person in the assembly saw the master's form. They heard only the

[14] Kyoto School philosophy: The philosophical system devised by a group of 20th century Japanese thinkers who all studied at Kyoto University in Japan. Their system brings together Western philosophies (e.g., existentialism, German idealism, etc.) and Eastern intellectual/spiritual traditions (e.g., Mahayana Buddhism, Zen Buddhism, non-dualism, etc.). Nishida Kitaro is often considered the originator of this system, though Nishitani Keiji's framework, which takes greater inspiration from Zen Buddhism, is widely considered a condensed version of Nishida's.

> sound of the Dharma. Among the gathering was Kanadeva. “Can you discern his form?” he asked the others. “Our eyes see nothing,” they answered. “Our ears hear nothing. Our minds discern nothing. Our bodies experience nothing.” Kanadeva said, “That itself is the form of the sage Nagarjuna Sonja manifesting the Buddha-nature. He is doing it to teach us. How do I know this? Because the form of formless samadhi is like the full moon. The meaning of the Buddha-nature is absolutely empty, clear and distinct.”
>
> (Dogen 2002, 77)

Being nondualists, however, Kyoto school philosophers further argue that nirvana is not truly nirvana until it is “turned” to rejoin the cycle of birth and death (i.e., everyday life), resulting in what they call samsara-*sive*-nirvana (“*sive*” meaning “as”) (Nishitani 1982, 179). This makes their views on free will and determinism so nuanced and ambiguous as to make any singular interpretation dubious. Besides this, Buddhist philosophers are notorious for their use of metaphors to vaguely explain key concepts, and nondualists are especially difficult to decipher. Because they believe that the ultimate truth lies beyond the reach of discursive thought, they often intentionally contradict themselves. Terms like “samsara-*sive*-nirvana,” “self as non-self,” and propositions like “each thing is itself in not being itself, and is not itself in being itself,” (149) only scratch the surface of the complex and intentionally contradictory ideas they regularly propose.

In contrast, Harris’ characterization of nondualism is mostly inspired by teachings from the Advaita Vedanta Hindu tradition by Ramana and Poonja-ji, and by teachings from the Tibetan Buddhist tradition by Dzogchen (Harris 2014). Though he emphasizes the experience of nondual awareness over and above its intellectual characterizations, the intellectual characterizations are entirely relevant in terms of deducing what the experience itself amounts to, and what the aims of meditative and spiritual practice are. Thus, it is rather unclear and highly debatable whether self-transcendence – the ultimate goal of meditative or spiritual practice – carries the implications regarding free will that Harris suggests it does. Though these considerations are intriguing and do have considerable impact on the discussion at hand, I do not wish to get stuck down this rabbit hole, since it is full of twists and turns that will take me too far afield. Suffice it to say, this brief aside was enough to demonstrate the questionable nature of Harris' claims regarding self-transcendence. At this point, I find it more impactful and manageable to address criticisms of the Libet experiment, the results of which Harris employs as empirical justification for his position.

A Critique of Libet-style Experiments

In 1985, Benjamin Libet and his team performed an experiment which supposedly showed that conscious decisions to "act now" are preceded by unconscious brain activity. Electroencephalography (EEG) was used to measure readiness potentials (RPs) in the brain, electromyography (EMG) to measure the timing of the participants' voluntary actions (wrist flicks in this case), and a cathode ray oscilloscope which functioned as a fast clock for measuring W times (i.e., the times at which participants become aware of their intention to act). The experiment was rather simple: 9 participants were seated in front of the clock, they were hooked up to the EEG and the EMG, then they were asked to flex their wrist whenever they felt like it and note the time when they had consciously decided to flex it on the fast clock. The timings recorded by participants were labelled as W times: "W" meaning "willing" or "wanting." It is important to note that participants were instructed not to pre-plan their voluntary movements to ensure that all recorded voluntary acts were spontaneous and endogenous (Libet 2004, 126). What Libet found was that RPs preceded recorded W times by an average of approximately 350-400ms (across 40 voluntary actions per participant) (134). These findings led him to believe that voluntary actions are not made consciously, since they are preceded by unconscious brain activity in the form of RPs (136). Upon obtaining these results, Libet then asked participants to plan to flex their wrist at a specified clock time, and then to cancel the action 100-200ms before the specified clock time, resulting in no wrist flexion (138). What he found in these cases was that a substantial RP developed some 1-2 seconds before the specified clock time, but that the RP then flattened in the 100-200ms window in which participants were instructed to veto that action (138). These findings led Libet to believe that though we do not consciously initiate our voluntary actions, we are able to consciously veto those actions (139).

Some, like Harris, interpret these results as proof that free will does not exist, since they suggest that all supposedly voluntary actions are preceded by unconscious brain activity. There are problems with this interpretation, however. First, this interpretation does not consider Libet's findings related to veto power, and as far as I know, Harris does not provide any criticism of Libet's experimental design, nor his interpretation of these findings, thus leading me to believe that in citing the Libet experiments without considering his ultimate conclusion, Harris is cherry-picking information that aligns with his pre-conceived deterministic framework. "As someone put it, Libet believed that although we don't have free will, we do have free won't" (Mele 2014, 12). There is a problem with Libet's own interpretation as well, however. His experimental design and interpretation of his own findings have themselves been called into question. These criticisms shed doubt upon the validity of his conclusions, regarding both the unconscious ground of voluntary actions, and the conscious ground of veto power. Ironically enough, Dennett is one among those who criticizes Libet's experimental design. Criticisms have also been provided by Alfred Mele, whose insights are particularly poignant given his participation in Libet's experiment (9). The Libet experiment, though well-known, is rather old, and there have been more recent attempts to reproduce his findings, which employ alternative experimental designs. It is also important to note that Libet-style experiments are not the only experiments cited by determinists as evidence that free will does not exist. Unfortunately, analyzing every one of these experiments is far beyond the scope of this thesis. The task I now turn to is that of recounting Dennett's and Mele's criticisms of Libet's experiment, providing an overview of a more recent Libet-style experiment, and demonstrating how the criticisms provided by Dennett and Mele apply to old and new experiments alike.

Dennett provides multiple criticisms of Libet's interpretation of his experimental results: One based on his invocation of the Cartesian Theatre to supposedly guarantee that participants are capable of accurately recording W times, one based on an alternative to Libet's Stalinesque[15] interpretation of the

[15] Stalinesque interpretation: That a secondary stimulus somehow prevents the conscious experience of a primary stimulus. These interpretations apply in cases where theorists attempt to decipher how stimuli can influence us without

results, and one based on the unnatural quality of the task itself. Regarding the first criticism, because the MDM suggests that there is no Cartesian Theatre, and that different perceptual information about the fast clock travel along different pathways in the brain which do not come together to form a single cohesive picture in consciousness, Dennett questions the accuracy of recorded W times:

> There is essentially continuous representation of the spot (representing it to be in various positions) in various different parts of the brain, starting at the retina and moving up through the visual system. The brightness of the spot is represented in some places at some times, its location in others, and its motion still in others. As the external spot moves, all these representations change, in an asynchronous and spatially distributed way. Where does "it all come together at an instant in consciousness"? Nowhere.
>
> (Dennett 1991, 165)

In effect, Dennett proposes that the MDM calls into question the participants' ability to accurately designate the moment they become conscious of their intention to act by using the fast-clock as a reference, since information about the clock is interpreted by different brain areas in competition with one another. He also criticizes Libet's interpretation of the experimental results, since he promotes what Dennett calls a Stalinesque interpretation where an Orwellian[16] interpretation is also possible. According to the Orwellian interpretation, also provided by Dennett, a conscious experience can be immediately forgotten or overwritten by a new conscious experience, leaving the original experience unrecorded (164). Libet does briefly consider the Orwellian interpretation, though he dismisses it since the Orwellian phenomenon is not empirically testable (Libet 2004, 66). I do not believe these objections are reasonable, however. The first relies on the notion that the Cartesian Theatre is an illusion – a highly debatable assumption despite the

our conscious awareness of them, otherwise known as memory illusions. The term itself was introduced by Dennett in *Consciousness Explained*.

[16] Orwellian interpretation: That a secondary stimulus immediately overwrites the conscious experience associated with a primary stimulus. This interpretation of memory illusions is as valid as the Stalinesque interpretation. The term itself was also introduced by Dennett in *Consciousness Explained*.

virtues of the MDM – and the second both cannot be empirically verified and does not take into account the fact that Libet calculated average W-times across 40 trials for each participant (127). Moreover, as far as the Orwellian interpretation is concerned, it is questionable whether a "conscious" experience that was never recordable by a subject can be considered a conscious experience at all. It is no surprise that Dennett should object to the validity of Libet's experiment on such grounds, considering his proposed method of phenomenal investigation treats the phenomenal worlds of participants as fictions. I already explained the issues with this method in previous sections, and given the preceding criticisms are grounded in that method, they do not constitute strong objections to Libet's interpretation of his experiment.

That said, Dennett's third criticism is much stronger than the first two, since he rightly points out the unnatural scenario participants were placed in during the experiment:

> […] the subject's task of determining where the spot was at some time in the subjective sequence is itself a voluntary task, and initiating it presumably takes some time. This is difficult not only because it is in competition with other concurrent projects, but also because it is unnatural – a conscious judgement of temporality of a sort that does not normally play a role in behaviour control, and hence has no natural meaning in the sequence.
>
> (Dennett 1991, 165)

This criticism casts doubt upon the generalizability of Libet's experimental results, and it parallels the criticisms offered by Mele, which I now analyze since he provides a more detailed critique on these grounds. Mele, having participated in the Libet experiment, points out the relevant distinction between arbitrary choice and intentional choice. Though some participants in the Libet experiment did preplan their wrist flexions, they were expected to report these preplanned volitional acts and keep them minimal, since Libet was concerned with recording the timings of arbitrary and spontaneous conscious decisions to act only (Libet 2004, 126). Mele points out that this fact invalidates Libet's generalized conclusion that no volitional acts are initiated consciously, on the grounds that many of the decisions we make in everyday life are not merely arbitrary, but preplanned and motivated by our own pre-established goals (Mele 2014, 14). To illustrate the distinction between arbitrary and intentional decisions, he uses the example of picking a jar of nuts while grocery shopping:

> If someone were to ask you why you picked up the jar you just now put in your shopping cart rather than any of the other jars on the shelf, you'd be in my situation when I'm asked why I said "now" when I did. "I don't know" would be an honest answer. Now, maybe you wouldn't say that free will is involved in picking up the jar of nuts; you might think that free will is too important to be involved in trivial tasks. But even if you think it is involved, free will might work very differently in this scenario rather than when weighing pros and cons and having to make a tough decision. You wouldn't want to generalize from Libet's findings to all decisions— including decisions made after a careful weighing of pros and cons. It's a huge leap to the conclusion that all decisions are made unconsciously. Maybe, when we consciously reason about what to do before we decide, we are much more likely to make our decisions consciously.
>
> (14)

In this example, Mele is drawing a distinction between the arbitrary decision of which jar of nuts to pick up, and the intentional decisions of whether to buy nuts in the first place, which kind of nut one wants to buy, which grocery store to buy them from, how much one is willing to spend, or even how many jars to pick up. This criticism alone is enough to cast doubt upon Libet's generalized conclusion, since it clearly shows that he had not accounted for every kind of decision participants could make. Besides this criticism, however, Mele also offers another strong objection related to Libet's potential conflation between preparation and intention as it relates to RPs. He argues the fact that RPs flatten when vetoing voluntary actions indicates the possibility that RPs do not coincide with intentions to act, but merely preparations to act regardless of whether the participant truly intends to flex their wrist or not (19). It is indeed questionable whether a participant instructed to cancel a prepared action ever truly intended to perform the act in the first place. This concern is further justified on the grounds that Libet instructed participants not to preplan their voluntary actions and to veto their voluntary actions. Not only do these instructions provide exogenous motivation, where Libet was seeking to measure intentional actions motivated endogenously (Libet 2004, 128), but they also play a role in fabricating an unnatural scenario. As per Mele's jar of nuts example: seldom-if-ever is it the case that we are instructed by a third party to arbitrarily pick a jar of nuts, or to veto the act of arbitrarily picking one.

Though I believe these criticisms are strong enough on their own to conclude that Libet's experimental results do not conclusively support his generalized interpretation of them, there is one further criticism that warrants attention. In an earlier section I alluded to the idea that Libet's insistence that a cognitive theory which takes full stock of the phenomenal features of consciousness is more desirable than the alternative was hypocritical. Given the preceding analysis of his experiment and interpretation of the results, it may already have become clear to you how this is the case. The problem, of course, is that his generalized conclusion is at odds with this idea, since we all possess a phenomenal experience of initiating our own voluntary actions, and his conclusion denies the validity of this experience. Ironically, Mele's theory that RPs measure preparations to act rather than intentions to act aligns more closely with the more desirable cognitive theory Libet calls for than his own interpretation does.

At this point, it has been established that the original Libet experiment performed in 1985 does not conclusively rule out the possibility of free will, but the original experiment performed by Libet is not the only experiment of this kind to be performed. In fact, the supposed success of Libet's original experiment and his introduction of veto power into discussions of free will has inspired multiple attempts at reproducing his results. Some of these attempts have made use of similar experimental designs, though the design of a recent experiment performed in 2016 by Matthias Schultze-Kraft *et. al* that reproduced Libet's original results is quite dissimilar from his original experiment. As in the original experiment, EEG was used to measure RPs and EMG was used to time the muscular contractions concordant with voluntary actions, but this is where the similarities come to a halt. Unlike the original Libet experiment, the 2016 experiment was a go-signal experiment. That is, instead of being sat in front of a fast clock, participants were instead sat in front of a computer screen functioning as a light which flashes either green (indicating "go") or red (indicating "stop"). Participants were also presented with a floor-mounted button, so EMG was used not to time wrist flexions, but ankle flexions. This experiment made use of a brain-computer interface (BCI) – a linear classifier capable of detecting the RPs of participants and controlling go/stop signals – and progressed through three stages instead of the two included in the original Libet experiment (Schultze-Kraft *et al.* 2016, 1080). I will now provide an overview of the experimental procedure, the results, and the interpretation of those results by the experimenters.

In stage one, 12 participants sat in front of the screen and were provided a go-signal. They were instructed to press the button any time after a 2 second interval following the appearance of the go-signal. Stop-signals were elicited randomly, so movements were not being predicted in this stage. Nevertheless, participants earned points by pressing the button while the go-signal was still active and lose points by pressing the button while the stop-signal was active. RP data collected from EEG recordings in stage one was then used to train the BCI to predict participant movements in the next two stages. Notably, two participants were removed from the experiment after stage one, since their RP amplitudes were too low for the BCI to predict their movements above chance level (1084). In stage two, the BCI was employed to predict participant movements in real-time. In effect, the same procedure would play out as in stage one, but the BCI would attempt to turn the go-signal to a stop-signal just before the participants pressed the button. In this stage, participants would earn points for pressing the button while the go-signal was active, and the BCI would earn points if they pressed the button while the stop-signal was active. Stage three played out the same way as stage two, the only difference being that after stage two was completed, participants were informed that their movements were being predicted by the BCI, and that they would have to move unpredictably in order to earn points in stage three (1080). Each participant performed an average of 326 trials in total across all three stages (1084). Four outcomes were identified for trials across the three stages of the experiment:

1. Missed Button Press Trial: The participant pressed the button while the go-signal was active, but no RP was detected. The participant earned a point. This outcome resulted from 66.5% of stage one trials, 31.5% of stage two trials, and 30.8% of stage three trials (1081).
2. Predicted Button Press Trial: The participant pressed the button within 1000ms of the stop-signal being active. The BCI earned a point. This outcome resulted from 1.2% of stage one trials, 19.5% of stage two trials, and 22.8% of stage three trials (1081).
3. Aborted Button Press Trial: The participant did not press the button within 1000ms of the stop-signal being active. EMG data indicated that they began to move but aborted mid-movement. Neither the participant nor the BCI earned a point. This outcome resulted from 2.2% of stage one trials, 15.2% of stage two trials, and 16.3% of stage three trials (1081).
4. Ambiguous Early Cancellation/False Alarm Trial: The participant did not press the button within 1000ms of the stop-signal being active. EMG data indicated that they never initiated a movement. This could have resulted either from a prepared action being terminated at an early stage, or false RP detection on the part of the BCI. These results were labelled “ambiguous” since there was no way to verify which of these alternatives was the case. Neither the participant nor the BCI earned a point. This outcome resulted from 30.1% of stage one trials, 33.5% of stage two trials, and 30% of stage three trials (1081).

Note the variation between rates of Missed Button Press Trials and Predicted Button Press Trials in stages two and three were marginal. These findings are corroborated by RP data, which shows no variation between the three stages of the experiment, despite participants having been informed that their movements were being predicted in stage three (1081). Experimenters also assessed how the timing of stop-signals related to the onset of movements indicated by EMG data. This assessment was done by analyzing the EMG and BCI data collected from Aborted Button Press Trials specifically. Experimenters found that hardly any participants moved when presented with stop-signals earlier than 200ms before EMG onset (1082), but that if the stop-signal was presented later than 200ms before EMG onset, then participants could not help initiating a movement, though they were still able to abort it (1083). Thus, they concluded that their experiment evinces Libet's claims regarding veto power, but we can only veto the initiation of movements if the veto occurs earlier than a 200ms window before initiating it. They dub this 200ms interval "the point of no return" (1084).

Now, my task is to demonstrate how Mele's criticisms of the original Libet experiment carry over to this new experiment. First, though participants scored fewer points in stages two and three, they were still able to earn points approximately 30% of the time, even without being informed that their movements were being predicted, as in the case of stage two trials. This that the RP data used by the BCI for predictions is not conducive to producing adequate predictions of participants' movements in general. Second, Mele's and Dennett's criticisms that Libet's original experiment placed participants in an unnatural scenario also carries over to this experiment, since seldom-if-ever is it the case that we are informed to make the timings of our actions unpredictable, though this information might crop up endogenously in the case we are playing competitive sports, for instance. The most poignant criticism that carries over, however, is Mele's criticism that RPs may not indicate intentions to move, but merely preparations to move regardless of underlying intentions. In fact, this experiment provides further support for this interpretation of what RPs truly indicate. This is because, unlike Libet, experimenters in this case did not instruct participants to avoid preplanning their actions, and they also had them participate in questionnaires to ascertain how their strategies changed from stage to stage: "When asked about their strategies during stages II and III, they reported 'not thinking about the movements' (5 of 10), 'pressing earlier' (4 of 10), or 'trying to be more spontaneous' (4 of 10)" (1082). These reports, in tandem with the non-variance of RPs detected in all three stages suggests that "thinking about moving," or intending to move, and preparing to move are distinct phenomena. Moreover, because RPs did not vary between the three stages of the experiment (1081), but intentions to move did, it must be that RPs measure only preparations to act.

One might be tempted, based on findings related to "the point of no return" that RPs still carry some significance when it comes to the notion that unconscious brain processes initiate, or at least restrict the outcome of voluntary actions, but the fact that even these movements can be aborted despite initiation suggests otherwise. Unfortunately, this study does not provide information about when the decision to abort these movements occurred. Did some participants preplan an aborted movement? Did some abort mid-movement? Without this information, interpretation of these results is difficult to say the least. Thus, neither the original Libet experiment, nor this more recent iteration provide any conclusive evidence that conscious intentions to act are initiated by unconscious brain processes, nor that intentions to act are expressed neuronally as RPs. Experiments of this kind therefore do not provide a solid foundation for hard determinism, as Harris believes they do.

Chapter 3

A Philosophical Approach

To recapitulate: thus far I have explained that the Chinese room thought experiment and Searle's biological naturalism pose no meaningful challenge to beliefs that entities like SPAUN are capable of conscious experience by demonstrating in detail the biological plausibility of its current iteration; I have laid out the philosophical underpinnings of the SPAUN research programme; I have critiqued the perspective of their most prominent exponent, Dan Dennett, and the hard-determinist perspective of Sam Harris; and I have shown that results of experiments in neuroscience which supposedly justify beliefs that free will is an illusion are truly inconclusive, motivating a philosophical approach to the issue. The task of this chapter is to construct a philosophical argument which demonstrates that dualistic interpretations of the phenomenon of free will are not unfounded, and that they provide a means of calling into question the conscious status of future entities like SPAUN. Rivers of ink have been spilled over the topic of free will and determinism over the last two thousand years or so, resulting in a plethora of arguments on each side. I do not intend to simply recite pre-existing arguments, but rather, produce a novel argument to encourage new discussions. To accomplish this, I will offer a demonstration which highlights an explanatory gap in the paradigm of modern physics which leaves room for dualistic interpretations.

Before I construct my argument, it is important that I clarify exactly what the argument is intended to demonstrate, what kind of free will and conception of self I intend to promote, and what assumptions motivate me to pursue this endeavour. Roughly, I intend to argue that there are reasons to believe that SPAUN-like entities, despite their biological plausibility, are F-zombies because the qualitative experience of freely deciding does not come free with brain properly structured brain activity regardless of the substrate. That is, I am going to show that because consciousness and the qualia that accompany it may well be explained as a meeting between the brain and some potentially non-physical entity or force, there is reason to believe that SPAUN is completely a physical system and therefore could not have done otherwise. This argument of course relies on the notion that we biological beings, unlike our machine counterparts, are not completely physically determined and therefore could have done otherwise. This leads me to the concept of self and kind of free will I promote. The self is, in this view, what could have done otherwise (or what can do otherwise) in the classical sense. To clarify, I am positing that the self can do otherwise potentially, and that this is its essence, but that this purpose is only fulfilled in cases where it is actualized. For example, the purpose of a hammer is to sink nails, but if no one ever picks up the hammer, its purpose is never fulfilled. This brand of free will I promote is not necessarily classical, insofar as it remains agnostic regarding the existence of souls or God and therefore has no ties to any religious or spiritual system. It more closely resembles the "ambitious" free will characterized by Mele. He distinguishes this kind of free will by its assumption of what he calls "deep openness," and he contrasts it with the "modest" free will resembling Dennett's compatibilistic view, and with the "magical" free will resembling the classical view (Mele 2014, 78-80). One of the primary focusses of the upcoming section will also be to establish exactly what this "deep openness" would have to be to secure a purpose for the self; that purpose being to do otherwise.

As for the assumptions underlying and motivating this argument, the issues plaguing Dennett's brand of compatibilism and his method of phenomenal investigation pointed out in the previous chapter clarify that my working assumption parallels Libet's sentiment that the most desirable of cognitive theories takes full stock of our phenomenal experiences, since otherwise we risk explaining away the supposedly obvious, leaving large explanatory gaps in our theories, and devaluing the first-person perspective. A secondary motivation is to establish the moral boundaries surrounding potential future iterations of SPAUN, or fully physically realized SPAUN-like entities. Considering the real possibility of developing such machines, and the motivation to do so, it behooves us to have our ducks in a row before they come about. With all disclaimers and clarifications out of the way, I will now present my demonstration and argument.

Temporal Zombies and The Simultaneity Problem

I previously discussed the possibilities surrounding two different types of zombies: Chalmers' P-zombies, and F-zombies. Chalmers brings up the possibility of P-zombies to establish a potential divide between first-person qualitative experiences and third-person cognitive behaviour, and to elucidate that cognitive science is still incapable of broaching the hard problem of consciousness. Chalmers inspires questions about the purpose of consciousness and the adequacy of naturalistic accounts of its existence like that provided by Dennett, but most of those interested in this debate are all too familiar with these two hypothetical entities. I wish to challenge physicalism in a new way, by exploring a specific phenomenon related to consciousness that the possibility of P-zombies does not already demonstrate. Thus, I now introduce what I call temporal zombies (T-zombies) as yet another hypothetical entity to serve as a demonstration highlighting a specific problem, which neither physics nor physicalism have succeeded in solving.

T-zombies have much in common with P-zombies, though unlike P-zombies, they do not technically constitute entities completely devoid of conscious qualia. Allow me to explain. Imagine that you are having a conversation with me at some arbitrary point in time, perhaps concerning our thoughts on *The Lord of the Rings* Trilogy directed by Peter Jackson. It would seem reasonable to assume that while we are having this conversation with one another, we are both conscious in that very moment together, simultaneously. When you say something to me, I am conscious of it and respond in turn; when I say something, you are conscious of it and respond in turn. Now, imagine that this is not truly the case. Imagine that we are not conscious in the same moment, but that the moment of my conscious experience is shifted so that it is, say, 2 hours ahead of your own. Let us assume that the hypothetical conversation in question lasts for 1 hour, and in your mind, we are in the middle of it. In my mind, then, the conversation ended 1 ½ hours ago.

Thus, while we are in the middle of the conversation, neither one of us is interacting with a conscious being, but a zombie – a T-zombie to be precise – devoid of the conscious qualitative experiences we assume each other possess in that moment, and yet we are both still conscious beings in our own frames of reference. That is, if you asked if I am conscious during the middle of the conversation (which you are now experiencing and which I experienced 1 ½ hours ago), I would say yes, and this would not be a lie, since the past is immutable, and 1½ hours ago when I said yes, I was. If I asked you if you are conscious at the end of the conversation (which you will experience in half an hour and which I experienced 1 hour ago), you would say yes too, and this would not be a lie either, since the future is predetermined, and you will be conscious once you experience that part of the conversation in half an hour. Neither one of us would be lying about the fact that we are conscious, and yet neither one of us would truly be present to the other at the time of acknowledgement. It is not that we are both conscious and not conscious in the same moment together, it is that we are both conscious in different moments, such that we only ever interact with each other's bodies. To clarify, what I mean to illustrate here is a scenario in which all bodies occupy the same shared physical timeline so they are all simultaneous with one another and move through real time at the same rate, but the conscious qualitative experiences of individuals occupy their own separate timelines which need not be simultaneous and need not move through subjective time at the same rate.

Let us further explore how this shifting in the moments of our respective consciousnesses might come about. Imagine that both of us, at some arbitrary point in time before our conversation, had both watched every film in *The Lord of the Rings* Trilogy in succession (I choose this series because of its popularity and extraordinary length). This of course must be the case, hypothetically speaking, since our hypothetical conversation concerns our thoughts on this very trilogy. Assume that we both watched the same version of this trilogy as well: the roughly 9-hour long theatrical version. Let us also pretend that you do not like high-fantasy or lengthy films and were coerced into watching by a friend; you find the trilogy boring. (I say "pretend" because you may like the trilogy or may have no friends) On the other hand, let us pretend that this is my favourite film trilogy, and I love every second I spend watching it. (I say "pretend" because though I enjoy it, I am not sure if it is my favourite)

Now, we are all well acquainted with the chronoceptive phenomenon of our subjective sense of time accelerating or decelerating depending on our engagement in whatever activity is at hand. When enjoying ourselves, time seems to fly by, and when not, it seems to drag on endlessly. Imagine, then, that these accelerations and decelerations of subjective time have a lasting effect on the moment of one's conscious experience. That is to say: imagine that though bodies all share a real physical timeline, minds do not, such that an acceleration in one's subjective sense of time relative to another's results in their minds accompanying their bodies at different times. For example: I love the trilogy, so the time I spend watching it flies by. After watching it, I check the clock time and note the time I spent watching felt like only 8 hours instead of 9. You hate the trilogy, so the time you spend watching it drags on. After watching it, you check the clock time and note the time you spent watching felt like 10 hours instead of 9. Perhaps this phenomenon accounts for the shifting of moments of consciousness. Subjectively speaking, I finished the trilogy in 8 hours, and you in 10, resulting in the 2-hour shift between our moments of consciousness.

All of that is patently absurd. I want to reiterate that though this prospect cannot be falsified and is certainly possible (according to specific interpretations in modern physics I cover in a later section), I do not believe it is truly the case. Furthermore, I do not intend to argue that SPAUN, if it ever were to achieve consciousness, would be a T-zombie either. Regardless of whether T-zombies truly exist or not, my main purpose for introducing them is merely to shed light on some of our core assumptions about the nature of consciousness, which often go undisclosed and uninvestigated. In fact, I am going to attempt to use the absurdity of T-zombies to my argumentative advantage. If you are anything like me, you would find the claim that T-zombies really exist deeply offensive for multiple reasons. Outlining these reasons will enable me to clarify and analyze certain intuitions we may have about the temporality of consciousness and physical reality:

1. If you believe you possess “ambitious” free will, this scenario is offensive because if my moment of consciousness is 2 hours ahead of your own, you are not free to have done or said otherwise during our conversation. I, being 2 hours ahead of you and having already experienced the conversation, know everything you said before you even had the chance to say it; what you are going to do and say is fated. Even if it happens to be that your moment of consciousness is ahead of mine, the likelihood that someone else’s moment is ahead of yours will still worry you. This will not bother you, however, if you are a hard determinist.
2. If you are an identity theorist or emergentist, this scenario is offensive because it assumes a temporal divide between the qualitative experiences that you believe come free with brain activity, and the brain activity itself. In other words, you may be disturbed by the idea that what brain activity in this moment along the supposedly shared physical timeline between our bodies measured by the clock is “special” for me – insofar as it is accompanied by qualia – is different from yours. That is, you may find it inconceivable that bodies all share one timeline, while minds do not.
3. If you are a physicalist, or perhaps a Leibnizian, this scenario is offensive because it suggests that there is no objective, physical, third-person sense of the now. That is, it challenges the intuition that the supposedly simultaneous moment of consciousness is determined by universal physical processes beyond subjective experience, namely, entropy or the second law of thermodynamics. It also inverts the standard intuition that clock-time is objective, and that our subjective sense of time is validated or invalidated by clock-time. In other words, time in this scenario is fundamentally a subjective phenomenon entailing a kind of “temporal solipsism.”

Indeed, these are all reasonable concerns. I wish to draw your attention closer, however, to the first and the third among them. The second concern, though reasonable as well, appears to depend on either the first or the third, since identity theory and emergentism do not necessarily posit the simultaneity of conscious moments between individuals, only that properly structured brain activity is accompanied by, or produces conscious experience. Whose brain activity we reference, and when, is arbitrary. If, as identity theorists, we are disturbed by the prospect of T-zombies, it must therefore be that we are disturbed by the challenges it poses to our intuition that we are simultaneously conscious in the present moment because it is the only moment in which we can be conscious, whether because we believe physical law has determined this to be the case, or because we believe we are free and the prospect of T-zombies denies us this belief. I dub the problem disclosed by the prospect of T-zombies as "The Simultaneity Problem." I am going to analyze which of the two aforementioned intuitions provides a more tenable solution to the problem, but before I do, it is important that I further clarify and distinguish what I mean by "the simultaneous moment of consciousness," since the astute may have already noted its resemblance to ideas discussed in the previous chapter.

One might be tempted to draw parallels between the prospect of T-zombies and the MDM to argue that this model might account for the hypothetical shifting between moments of consciousness. Such a comparison, however, would be mistaking the moment *of consciousness* for the moment one becomes *conscious of* some phenomenon. Recall Dennett's criticism of the Libet experiment related to the fact that different information about the fast clock travels through different pathways in the brain which process information at different rates, based on the MDM. One might believe that the minor differences in rates of information processing of different brain areas might provide an account for shifts in the conscious moment between individual cognisors. This is not the case. It is reasonable to assume we all agree that our belief that the consciousness of others is simultaneous with our own should be held regardless of the contents of other minds. For instance, while you are reading this, you are conscious of the words on this page and their meaning, and even if I am, in this very same moment, cooking dinner at home, my consciousness is still simultaneous with yours despite the difference in content.

Applying this same logic to Dennett's MDM and the Libet experiment, it becomes clear that there is a distinction to be made between the rate at which contents enter consciousness, and the rate at which the conscious moment itself, hypothetically empty of all contents, moves through real time. In other words, even if we were to imagine a case where we were both participants in a Libet experiment, and one of our brains could process information at a marginally faster rate – hypothetically resulting in more accurate or swifter recordings on the fast clock – we would still agree that despite the differences in accuracy and speed, our conscious moments were simultaneous, and differed only in their contents from moment to moment. The MDM does not imply, and therefore does not account for, any shifting in the conscious moments between individuals, though it may imply diverse rates in terms of the acquisition of conscious contents. Thus, despite Dennett's insistence that the Cartesian Theatre is incoherent and illusory (a debatable claim in itself) there is yet to be any meaningful challenge to the prospect of a "Cartesian Date," so the door is not shut on dualism just yet. I take it this prospect is one worth protecting at all costs.

The Problem of Now

Now that I have adequately defined the problem, and clarified key terms, I move on to analyzing whether and how physicalism deals with it. In other words, the question to be answered is whether there is an adequately established physical explanation for the simultaneity of consciousness. The purpose of this section will be to demonstrate that physics provides no such explanation, and that this fact justifies an explanation grounded in "ambitious" free will. I am aware this is a bold claim, so before I get into the analysis, I wish to acknowledge that none of what I am about to say is intended to deny the virtues of physical accounts, nor the possibility that an adequate physical account might one day be provided. The purpose is merely to point out the existence of an explanatory gap that leaves room for dualistic accounts and solutions to the problem.

Interestingly, a problem like The Simultaneity Problem has already been acknowledged within the physics community, by none other than Albert Einstein himself. It is important to note that not much ink has been spilled over this problem in the physics community, as it is not addressed in any of Einstein's physical theories. His theories of special and general relativity rather concern physical systems and clock-times, but there is evidence suggesting that he had concern for a problem related to our subjective sense of time, and that he discussed it with at least one of his known associates, Rudolph Carnap:

> Once Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future, but that this important difference does not and cannot occur within physics. That this experience cannot be grasped by science seemed to him a matter of painful but inevitable resignation. I remarked that all that occurs objectively can be described in science; on the one hand the temporal sequence of events is described in physics; and, on the other hand, the peculiarities of man's experiences with respect to time, including his different attitude towards past, present, and future, can be described and (in principle) explained in psychology. But Einstein thought that these scientific descriptions cannot possibly satisfy our human needs; that there is something essential about the Now which is just outside of the realm of science. We both agreed that this was not a question of a defect for which science could be blamed, as Bergson thought. I did not wish to press the point, because I wanted primarily to understand his personal attitude to the problem rather than to clarify the theoretical situation. But I definitely had the impression that Einstein's thinking on this point involved a lack of distinction between experience and knowledge. Since science in principle can say all that can be said, there is no unanswerable question left. But though there is no theoretical question left, there is still the common human emotional experience, which is sometimes disturbing for special psychological reasons.
>
> (Carnap 1963, 37)

The modern physics community has rarely addressed this issue directly, and this may be due to issues of interpretation, or the fact that this second-hand information provided by Carnap is the only account of the problem that exists. Exploring the few existing interpretations is therefore both manageable and necessary, since those who have addressed this directly, namely, Carnap, Sabine Hossenfelder have typically interpreted the problem in a particular way that minimizes its relevance within physics. I maintain that both Carnap’s and Hossenfelder’s interpretations of the problem are not sufficient, and that the prospect of T-zombies introduced in the previous section provides clarity in this regard. It is important to note that

Einstein's concern is related to the implications and his own theories, since his introduction of spacetime led to the construction of eternalist block universe models which posit the equal existence or "presence" of past, present, and future events, treating every moment in time as equally special (or un-special) and thus our experience of the flow of time as an illusion (Ellis 2006, 1). I outline this problem in greater detail in the upcoming section and bring it up here only to clarify the motivation behind the discussion between Einstein and Carnap recounted in the preceding passage. Before doing that, however, I must show that Carnap and Hossenfelder do not provide adequate interpretations of the problem as Einstein poses it and offer one that does not minimize his concern.

To start, from the preceding quote it is clear that Carnap did not understand the problem Einstein meant to identify. He says: "I definitely had the impression that Einstein's thinking on this point involved a lack of distinction between experience and knowledge" (37), when Einstein was evidently making a distinction between the two. As Carnap recounts, Einstein explained that "the *experience* of Now means something special for man" (37), and that "this experience cannot be grasped by science" (37). That is, Einstein was making a distinction between the *knowledge* gleaned from the third-person perspective of the physical and psychological sciences, and the *experience* gleaned from the first-person perspective of consciousness. It appears the disagreement between them was not due to a lack of distinction between knowledge and experience, but due to differing beliefs regarding whether knowledge explains experience. Carnap is at least correct when he says that the temporal sequence of events can be described in physics, and that the differing attitudes toward the past, present, and future can in principle be described psychologically, but he drastically minimizes the problem as Einstein presented it. Einstein was not questioning our ability to describe when the Now is, or when the past and future are, he was questioning our ability to explain *why the experience of Now is when it is*. It is important to note that this is my own interpretation of what Einstein was saying. I choose to interpret him this way for one simple reason: he claims that "this important difference does not and cannot occur within physics" (37), and the question of why the experience of Now is when it is points to phenomenon that physics indeed struggles to explain. I also take it that if any figure in the history of modern physics had the authourity to say what could or could not be explained scientifically, it would be one with as great an intellect as Einstein. That is, I interpret the problem this way because it is charitable.

This interpretation of the problem of Now allows me to draw parallels between it and The Simultaneity Problem, though the problem of Now has a much broader scope. The Simultaneity Problem calls for an explanation of the simultaneity of conscious moments between individuals, whereas the problem of Now calls for an explanation of why, assuming the conscious moments of all sentient beings are simultaneous, the moment itself is this point in time. For instance: why is the Now we all experience not 2 hours ago, or 2 hours ahead? This is the question Einstein was asking, and he seemed to believe, based on Carnap's account, that the sciences are incapable of answering it. Assuming Einstein was correct about this, the incapacity to solve the problem of Now entails the incapacity to solve the Simultaneity Problem. After all, if we cannot explain why the experience of Now is when it is, then how can we explain why your Now and my Now are simultaneous? I am going to argue that Einstein was right to be concerned, at least insofar as the sciences have yet to solve these problems, but before doing so, I will first address Hossenfelder's treatment of the problem in a 2014 post on her personal blog, since it is the only direct attempt at solving the problem that I have been able to find.

To start, Hossenfelder argues that introducing the problem of Now constituted Einstein's greatest intellectual blunder. Worse even than introducing the cosmological constant or stating his conviction that God does not throw dice (Hossenfelder 2014). She then goes on to offer her interpretation of the problem as presented by Einstein and recounted by Carnap:

> The problem is often presented like this. Most of us experience a present moment, which is a special moment in time, unlike the past and unlike the future. If you write down the equations governing the motion of some particle through space, then this particle is described, mathematically, by a function. In the simplest case this is a curve in space-time, meaning the function is a map from the real numbers to a four-dimensional manifold. The particle changes its location with time. But

> regardless of whether you use an external definition of time (some coordinate system) or an internal definition (such as the length of the curve), every single instant on that curve is just some point in space-time. Which one, then, is "now"?
>
> (Hossenfelder 2014)

You may have already noticed that Hossenfelder's interpretation drastically minimizes the import of Einstein's concern. He was not asking "which one is Now?" He was asking "why Now?" Keep this distinction in mind as I further describe Hossenfelder's proposed solution to the problem as you will find this misinterpretation renders it insufficient. She goes on to explain that the problem does not call into question the potential for a mathematical explanation of our differing attitudes toward the past, present, and future (Hossenfelder 2014). To justify this, she argues that our differing attitudes can be accounted for by the fact that we have memory, and to prove this, she offers a mathematical formalization:

> If we want to describe systems with memory we need at the very least two time parameters: t to parameterize the location of the particle and τ to parameterize the strength of memory of other times depending on its present location. This means there is a function $f(t,\tau)$ that encodes how strong is the memory of time τ at moment t. You need, in other words, at the very least a two-point function, a plain particle trajectory will not do. That we experience a "now" means that the strength of memory peaks when both time parameters are identical, i.e., $t - \tau = 0$. That we do not have any memory of the future means that the function vanishes when $\tau > t$. For the past it must decay somehow, but the details don't matter. This construction is already sufficient to explain why we have the subjective experience of the present moment being special. And it wasn't that difficult, was it?
>
> (Hossenfelder 2014)

No, it was not that difficult. Unfortunately, that is only the case because this does not solve the problem as it was posed. Recall Einstein's remark that "these scientific descriptions cannot possibly satisfy our human needs," because "there is something essential about the Now which is just outside of the realm of science" (Carnap 1963, 37). Why would we not be satisfied by this? One reason is because this gives no explanation for why the variable "t" in Hossenfelder's formulation contains whatever value it happens to contain. That is, unlike the mathematical quantity "t," which is a variable that can contain any arbitrarily assigned value from the third-person perspective of science, the conscious moment of subjective experience is invariably Now; eternally present. Einstein's concern, then, could be due to the fact that because the sciences depend on a third-person perspective, they cannot explain why "t" contains one value and not another. On this interpretation, Einstein was not merely contemplating the laws of physics; he was contemplating the law of the laws of physics. I must clarify that I agree with both Einstein's and Carnap's statements that these issues do not result from some defect for which the sciences can be blamed. It may be the case, after all, that Einstein was wrong about this, and that physics will one day provide explanations for such things. Nevertheless, it may also be the case that the third-person perspective upon which the sciences depend places fundamental restrictions on the kind of explanations and knowledge they offer.

Now, I have shown that Hossenfelder, despite her best efforts, only provides a mathematical formulation explaining and describing our psychological attitudes toward the past, present, and future, which misses the point Einstein intended to make. In her closing statements on this issue, she mentions that: "In the above construction all moments are special in the same way, but in every moment that very moment is perceived as special. This is perfectly compatible with both our experience and the block universe of general relativity. So Einstein should not have worried" (Hossenfelder 2014). This is a very interesting declaration, for a couple reasons. First, it clarifies that what Hossenfelder has truly offered with this supposed solution is exactly the kind of psychological explanation for, and description of, our differing attitudes to past, present, and future that Carnap believed possible. Clearly, he was right about this, as Hossenfelder proves. Second, it introduces block universes into the discussion, and drives home the point that mathematical constructions treat all moments as special in the same way. This latter point highlights the problem Einstein was getting at: in conscious experience, not all moments are special in the same way, only the present moment is. As for the former point regarding block universes, Hossenfelder unfortunately does not specify what kind of block universe she is referencing. It is worth establishing what kind of block universe is truly compatible with our experience, since doing so will allow me to further characterize the explanatory gap and disagreements in physics, establish the kind of soft-determinism I endorse, and illustrate the role “ambitious” free will might play in filling the explanatory gap.

Block Universes

Block universe models have been a mainstay in theoretical physics since Einstein introduced the theories of special and general relativity. They specifically developed as a consequence of the notion of spacetime present within his theories (Ellis 2006, 1). These models contrast with classical theories and philosophies of time which promote the real distinction between past, present, and future, in that, much like Hossenfelder’s description of them, they do not take any particular moment in time to be in any way “special.” That is, classical philosophies of time typically promote either growing/evolving block universe models which view past and present as real but the future as unreal or empty (a.k.a. possibilism), or non-block models which only view the present as real (a.k.a. presentism). Since Einstein, however, modern physics tends to promote standard block universes (1), which take past, present, and future to be equally real (a.k.a. eternalism) (Savitt 2021). What really makes the difference between the classical and modern view are beliefs regarding the status of the future and the flow of time we all experience.

Eternalists would have us believe that all future events are predetermined, since in some sense they “already exist,” or are “equally as present” as any past or present events (Ellis 2006, 1). As professor of philosophy at the University of Sydney and philosopher of time Kristie Miller puts it: “Eternalists view objects that exist (and events that occur) at other times as being analogous to objects that exist (and events that occur) at other places. Just as Singapore exists, despite not existing here in Sydney, so too dinosaurs exist, despite not existing now” (Miller 2013, 346). For simplicity’s sake, one might describe an eternalist block universe model as a complete timeline comprising every event at every spatial point in the known universe, regardless of any single event's relation to the present moment. Theoretically speaking, one might conceive of an eternalist block universe beginning with pre-inflationary quantum fluctuations that resulted in the big bang and ending with proton decay and the heat death of the universe. The eternalist hypothesis that "The universe just is: a fixed spacetime block" (Ellis 2006, 1), also gives rise to the belief that our conscious experience of the flow of time is an illusion (1). Note the similarity between this claim and those of Dennett and Harris.

It is important to note that though eternalists believe the flow of time as we experience it is illusory, they still believe in objective temporal relations between positions on the fixed spacetime block. That is, a particular moment on the block can still be in a relationship of simultaneity, earlier than, or later than with other events on the block, though the present moment in which we assume all our moments of consciousness are simultaneous has no privileged position in the block (Miller 2013, 346). A physicalist might be tempted to argue that because the order of temporal events is still accounted for in eternalist block universe models, T-zombies are not truly compatible with them, since they seem to violate this ordering. This argument relies, however, on the assumption of simultaneity between individual moments of consciousness and the notion that this simultaneity is physically determined, neither of which are necessarily linked to eternalism. T-zombies still abide by the relations of earlier than and later than posited by eternalists, because a T-zombie's conscious experience is not simultaneous, but earlier than or later than the conscious experience of one interacting with it, meaning T-zombies do not violate the order of temporal events posited by eternalists, and they are indeed compatible with eternalist models. In essence, the hypothetical existence of T-zombies relies on the possible invertibility of our typical understanding that individuals can be conscious at the same time in different places: it could also be that individuals are conscious at different times in the same place.

Another point to consider regarding eternalism is that a version of it exists which attempts to reconcile the model with the notion of a privileged present moment, called the moving spotlight view

(Miller 2013, 347). This view synthesizes the eternalist ontological thesis[17] (EOT) with the dynamical thesis[18] (DT) typically characterizing presentism, whereas typical eternalism rejects DT in favour of the static thesis[19] (ST); (347). Then again, no clear explanation is given for what this "spotlight" is, where it comes from, or how it comes about. That is, the "spotlight" is assumed by reference to one's present conscious experience (353), and the assumption that the conscious experiences of others are simultaneous with one's own. Moreover, the DT characterizing the moving spotlight view can only be reconciled with the theory of special relativity (STR) by positing that the present hyper-plane[20] is metaphysically privileged, since "According to STR there is no privileged hyper-plane" (353). In other words, it does not follow from STR that the objective present is physically determined, but metaphysically determined. In fact, "it is no part of our best theories in the physical sciences that any single hyper-plane is privileged. Empirically, if you will, all planes are equal: no experiment could reveal one hyper-plane to be privileged" (353). This fact aligns with Einstein's contention that the important difference between present events, and past and future events does not and cannot occur within physics (Carnap 1963, 37).

This is all to drive home the point that the moving spotlight view can only maintain DT by reference to the flow of time we all experience, and that this experience is evidence that the present is at least metaphysically privileged. Typical eternalists instead deny the experience of the flow of time as illusory, and thus their view seems to follow directly from STR. It is also worth noting that eternalists have also recently begun considering the possibility of retrocausation: that future events can determine present and past events within the block (Adlam 2022). This may strike you as a ridiculous consideration, but teleological precursors to this idea date back to Aristotle, who posited his prime mover as a final cause of all motion in the universe. It is also worth noting that the physics community is currently divided on this

[17] The Eternalist Ontological Thesis: That past, present, and future times and events exist.
[18] The Dynamical Thesis: That the present moves, and which moment is the present moment changes.
[19] The Static Thesis: That the present does not move, and which moment is the present moment does not change.
[20] Hyper-plane: A subspace whose dimension is one less than that of its ambient space. In the context of block universes, which are four-dimensional spaces, a hyper-plane is a three-dimensional subspace representing a moment in time on the block. In simpler terms, if representing a four-dimensional block in three-dimensions, a hyper-plane may be regarded as a two-dimensional "slice" of the block.

issue, both on the supposed reversibility of the flow of time, and whether retrocausation can be considered the same phenomenon as causation at all (Spekkens 2022).

Coming back to Hossenfelder's treatment of the problem of Now: given that there are two kinds of block models (eternalist and growing/evolving block models) available for consideration, and that growing/evolving blocks are designated by these terms, Hossenfelder's use of the standard and unqualified term "block universe," in tandem with her inadequate interpretation and treatment of the problem of Now, leads me to surmise that she was referring to the eternalist block universe model. This presents a problem, however, in that while she is correct that eternalist models are compatible with her mathematical explanation of our differing attitudes regarding past, present, and future, the problem Einstein was really getting at concerning the seemingly impossible task of explaining why our experience of the Now is when it is comes as a direct result of employing the very eternalist models his theories inspired, as evinced by the fact that it follows from STR that no hyper-plane is privileged. Contrary to her claim, then, that "Einstein should not have worried," his worries regarding eternalism's explanatory disadvantages have since been echoed by George F.R. Ellis: professor of complex systems and mathematics at the University of Cape Town. In a 2006 paper, Ellis argued that eternalists block universe models are unrealistic and should be replaced with more realistic evolving block universe models (Ellis 2006). His justification for this goes beyond the mere fact that "the present (`now') is not usually even denoted in the diagram" (1). He further argues that eternalism "does not take complex physics or biology seriously; and they do indeed exist in the real universe. The irreversible flow of time is one of the dominant features of biology, as well as of the physics of complex interactions and indeed our own human experience" (5).

While Ellis provides ample justification for his argument regarding the irreversibility of the flow of time as it relates to biology and complex physical interactions, I have not the expertise nor the remaining space to lay them out in detail. That said, it stands to reason that if eternalist block universe models treat every moment in time as equally special (or un-special), equally present, and treat the unidirectional flow of time as an illusion, an eternalist has no reason to object to the claim that T-zombies actually exist. In the eternalist picture, there is nothing special about the present, and therefore nothing special about brain activity occurring in the present, unless one adopts the moving spotlight view, in which case the question of what it is that metaphysically privileges the present moment is left unanswered. Moreover, the typical eternalist block universe picture accommodates T-zombies so well that even retrocausality is compatible with them. Recall the hypothetical scenario where we are both T-zombies and my moment of consciousness is ahead of yours' by 2 hours. Could not the fact that I would know everything you did and said during our hypothetical conversation play a role in retrocausally determining what you are going to do and say? After all, during the conversation, which is in my past, some of your actions would have come in direct response to mine. I think we can all agree that the absurdity of T-zombies entails that any picture of physical reality that accommodates them is untenable. In other words, because eternalism is compatible with the absurd existence of T-zombies, it cannot hope to solve The Simultaneity Problem, or the problem of Now. It must therefore be defective.

At this point, I wish to highlight Ellis' claim that eternalism does not take the human experience seriously, since it parallels Libet's argument that cognitive theories which take phenomenal experiences seriously are more desirable than their counterparts, which I have demonstrated attempt to explain them away; labelling them as "illusions." It also highlights that since the moving spotlight view does take the human experience seriously, it is a more tenable view, despite the invocation of the metaphysical. On a similar front, Ellis remarks that "Human intentionality underlies the unpredictable functioning of the mechanisms (motors, computers, etc.) […] as they would be the result of human agency (implied by their supposed existence as designed objects)" (13). This argument bears directly on the current discussion of whether the complexity and unpredictability of a designed object, like SPAUN, warrants the ascription of phenomenal states, something I am going to argue we should be wary of doing. This brings me back around to the task of developing an alternative view of the purpose of the self and its conscious experience to those I have thus far critiqued.

The Self Reappears

Taking everything discussed thus far into consideration, the task ahead of me is that of securing a purpose for the self which depends on a dualistic interpretation, for the purpose of calling into question SPAUN's innate capacity for selfhood, qualitative experiences, and free will, given its biological plausibility. My argument, plainly speaking, is that contrary to Dennett's and Harris' views on the subject, the self can do otherwise regardless of physical constraints placed on the brain, and that this is its purpose. For this to be the case, it must be either that the self is itself non-physical, or the result of an interaction between the brain and some mysterious and potentially non-physical entity/force. In other words, the door is still open for dualist interpretations of Being, since physicalism relies on interpretations of physical reality and theories of time which leave explanatory gaps that may never be filled. How does all this relate to the topic of whether SPAUN is conscious or free? Well, if dualism is true, and we can do otherwise, which we know from first-hand experience, there is reason to believe that SPAUN cannot do otherwise like we can, since function alone does not guarantee conscious states.

Ellis has provided me the necessary tools for describing a softer determinism than Dennett's, which might secure for the self its purpose and provide preliminary solutions for the problem of Now and The Simultaneity Problem. I say "preliminary" to emphasize that doors are also open for physical solutions to these problem, and I have only succeeded in demonstrating that none have yet passed through them. Perhaps I will be forced to one day eat my proceeding words. Given Ellis' characterization of eternalist block universe models, it is safe to assume that they imply a fatalistic, or hard-deterministic interpretation of physical causation in line with Harris' perspective. Ellis' criticisms of eternalism highlight its explanatory disadvantages with regards to the problem of Now and The Simultaneity Problem, thereby illustrating the shaky ground upon which Harris' position on free will stands. I therefore find it reasonable – considering also the vast difference between our viewpoints – to spend my efforts examining Dennett's position, since it is compatible with the more realistic evolving block universe model Ellis proposes.

To clarify, though I take issue with Dennett's philosophy on multiple fronts that I previously discussed at length, he and I ultimately disagree about one key matter: whether we can do otherwise. We agree wholeheartedly, however, on the fact that indeterminism cannot provide any semblance of free will, hence why I have chosen to entirely disregard presentism, since it is equally compatible with indeterminism and determinism. That said, Dennett's brand of soft determinism and his naturalism both appear, on the surface, to be compatible with the evolving block universe model as described by Ellis. After all, this model proposes that past events are as real as present events, leaving only the future undetermined, all of which is consistent with evolution theory. Problems arise, however, when we consider Dennett's claim that we cannot do otherwise. The problem, of course, is that this implies the future is determined by the past, a notion inconsistent with evolving block universe models, since without being able to do otherwise, there seems to be no reason why the collective moment of consciousness is now. In other words, Dennett's soft determinism too easily slips into hard determinism. I have already shown how insufficient Dennett's defenses against the criticisms of hard determinists are in the previous chapter, so I am not going to retread that ground. I only wish to highlight how the concern for whether we can do otherwise is directly related to our conceptions of mental and physical causation.

One may have already discerned that evolutionary block universe models are also compatible with a particular brand of hard determinism, one in which the future is only epistemically indetermined, rather than ontologically so. This is further evinced by the fact that evolutionary models are compatible with Dennett's claim that we cannot do otherwise. While this interpretation of evolutionary block universe models does provide a solution for the problem of Now, insofar as the present occupies a privileged and determinable position at the end of the block (Miller 2013, 353), it does not provide an adequate solution to The Simultaneity Problem. In other words, a hard-deterministic, physicalist interpretation of evolutionary models also accommodates the existence of T-zombies, since it cannot explain why *everyone's* moment of consciousness accompanies the physically determined present moment. Allow me to demonstrate. Extend the earlier hypothetical case where my moment of consciousness is 2 hours ahead of yours' to all sentient beings, so that everyone has a different moment of consciousness, and we are all T-zombies. It stands to reason that if this were the case, only one, or a handful of individuals would ever find their moment of conscious concurrent with the physically determined present moment, the rest of us being left behind. The preceding argument is essentially a revamping of what Miller calls "the epistemic challenge" (359) to evolving/growing block models:

> Consider Julius Caesar at the moment at which he crosses the Rubicon. There was a time when Caesar was in the objective present. Thereafter, he has been in the objective past. Suppose that in 60 BC, when Caesar is having a chat to Cicero about Pompey, Cicero asks Caesar whether either or both or them are located in the present. There is a moment – a durationless instant – at which, were Caesar to (very rapidly) answer "yes", he would speak the truth: namely when the relevant three- dimensional slice upon which he and Cicero are located, is at the very edge of the block. Thereafter, Caesar would be wrong to answer "yes". Cicero then puts it to Caesar that since there are either an infinite number (if time is continuous) or a very large finite number (if time is discrete) of locations in the four-dimensional block that are in the objective past, and only one instant that is in the objective present, that he and Caesar ought to think it far more probable that each of them is in the objective past (Bourne 2002; Braddon-Mitchell 2004; Merricks 2006). The problem to which Cicero alludes is that in a growing-block world, it does not seem possible to determine whether one is located in the objective present or the objective past, and given this, one should conclude that almost certainly one is in the objective past.
>
> (358)

This challenge is typically overcome by defenders of evolving/growing block universe models by arguing that "it is necessary for the existence of phenomenology at a time t, that t is the objectively present slice" (359). The interesting thing about this argument is that it proposes the existence of phenomenal states in the present is in some way dependent on the fact that there is no hyper-plane beyond the present hyper-plane, and vice-versa (359). The question to be answered at this point is: "why must phenomenal experiences coincide with the present hyper-plane at the end of the evolving/growing block?" I previously mentioned that a physicalist would find the prospect that T-zombies really exist to be offensive because they appear violate the second law of thermodynamics that establishes the concept of entropy: the tendency for physical systems to increase in disorder, randomness, or uncertainty over time. A physicalist may take entropy to be a physical, measurable phenomenon which determines that the present moment is when it is, and that phenomenal experiences are somehow linked to this physical process. There are two problems with this argument, however. First, the universality of the second law of thermodynamics has been challenged on both physical and computational grounds (Hemmo *et al.* 2020; Cápek *et al.* 2011). Second, it does not necessarily follow from the fact that the present moment is defined as the moment in time with the highest measurable entropy that phenomenal experiences must accompany the present. In other words, how physics determines that phenomenal experiences should only accompany the point in time with the highest degree of entropy seems unanswerable.

What I offer instead, then, is the argument that selves possessing phenomenal experiences find themselves only in the present because the self performs some function that can only be accomplished in the present moment. Although evolving block universe models can pinpoint the present moment, unless every conscious individual can do otherwise, and free will is not an illusion, no indisputable physical reason can be given for why everyone finds themselves in the present moment identified at the end of the block, since it follows from our modern understanding of physics that there may be nothing special about present brain activity when compared to past brain activity, or even future brain activity for that matter. This is to say, that though there are interpretations of the physics which a physicalist may advance to solve The Simultaneity Problem, these interpretations are not set in stone, since competing interpretations are possible, showing that none of them necessarily follow from the physics itself.

Thus, what I propose is a softer determinism which incorporates the interaction between the physical and the non-physical to guarantee the simultaneity of all conscious experience. As Ramachandran points out, the self and qualitative experiences go hand in hand, since there can be no experience without experiencer, and no experiencer without experience (Ramachandran 2011). It follows from this, that because qualitative experiences, including the experience of freely deciding, are non-physical, the self must either also be non-physical, or at least result from the interaction between the physical and some mysterious non-physical entity/force, otherwise there would be no interaction between the self and qualia. To accommodate this non-physicality and our ability to do otherwise, I introduce an adaptation to the evolutionary block universe model proposed by Ellis, the adaptation being a breach, or rupture, between past and present resulting from the influence of the non-physical. In other words, whereas the evolutionary model presented by Ellis posits that the past *determines* the present, and the future is empty, I posit that the past *influences* the present, and that the present *influences* the future. In this model, the present is taken to be the moment of consciousness in which the non-physical meets with the causal influence of past events to also causally influence the future. Ellis, for instance, remarks that "The present is more real than the undetermined future, in that it is where action is now taking place: it is where the uncertain future becomes the immutable past" (Ellis 2006, 6). Indeed, this follows from his proposed evolutionary model, but as I previously demonstrated, without the breach, or rupture between past and present resulting from the influence of the non-physical, the model can provide no adequate solution for The Simultaneity Problem. So, to answer the question of what is it that must be "deeply open" for "ambitious" free will to be possible: it must be the present.

Before concluding, it is important that I address how this view provides a preliminary solution for the problem of Now, and how that solution compares to that proposed by physicalists. I must admit, before sharing, that the solution is not stellar, though it is at least as adequate as the alternatives proposed by the physics community. The solution is this: the present is when it is because it is the only time at which conscious beings are free to do otherwise, and all conscious beings are in fact free to do so. In simpler terms, we find ourselves in the present because the influence of the non-physical presents itself only now. You will find that this reasoning is circular, and also begs the question, so may choose to repudiate it on those grounds. The only further defenses I can offer in this case are that, unlike the previous viewpoints I have critiqued in this paper, which provide no adequate solution for the problem of Now, nor The Simultaneity Problem, the view I am promoting at least solves one of the problems, and its solution to the other, though unsatisfactory, is no more so than proposed alternatives. After all, there is as much reason to be believe that the "law of laws" Einstein was contemplating are the result of unknown non-physical factors as they are the result of unknown physical factors. The physicalist notion that the ordering of events is physically determined (time evolution theory) is grounded by our phenomenal experiences, or by the assumption that identity theory is valid. The former justification is based on an intuition no more cogent than dualistic alternatives, and the latter justification is circular. Moreover, the view I propose has the benefits of taking phenomenal experiences seriously and maintaining the acceptability of the criminal justice system.

To briefly summarize: I have demonstrated SPAUN's biological plausibility and laid out the philosophical underpinnings of the theories that guide its modelling, I have critiqued philosophical perspectives which might unduly ascribe consciousness to SPAUN, and I have raised and solved The Problem of Simultaneity in an effort to demonstrate what might distinguish us from SPAUN. In effect, I have argued that because the door for dualistic interpretations of Being is still open, and the non-physical is not well understood (and perhaps never will be), we should be wary of ascribing conscious experiences to non-biological beings solely based on their ability to simulate cognitive behaviours. It may be that the non-physical does not interact with the non-biological as it does with the biological, which is to say, the underlying physical substrate of the system in question, or some other unknown factor, may play a definitive role in establishing what "internal" states or "cosmic" functions the system performs.

References

Adlam, Emily. "Two roads to retrocausality." *Synthese 200.5*, 2022.

Aquinas, Thomas. “Summa Theologica.” Translated by Fathers of the English Dominican Province, Benzinger Bros ed., *Public Domain*, 1947.

Aristotle. “Metaphysics.” In The Basic Works of Aristotle. Edited by Richard McKeon and translated by W.D. Ross. *Random House, Inc.*, 1941: 689-926.

Aristotle. “Physics.” In The Basic Works of Aristotle. Edited by Richard McKeon and translated by R.K. Gaye and R.P. Hardie. *Random House, Inc*., 1941: 218-394.

Bekolay, Bergstra, Hunsberger, DeWolf, Stewart, Rasmussen, Choo, Voelker & Eliasmith. “Nengo: a Python tool for building large-scale functional brain models.” *Frontiers in Neuroinformatics 7*, 2014.

Block, Ned. "Troubles with functionalism." The Language and Thought Series. *Harvard University Press*, 1980: 268-306.

Cápek, Vladislav, and Daniel P. Sheehan. “Challenges to the second law of thermodynamics.” *Dordrecht: Springer*, 2005.

Carnap, Rudolph. “The Philosophy of Rudolph Carnap.” Edited by Paul Arthur Schlipp. *Open Court Publishing Company*, 1963.

Chalmers, David J. “The conscious mind: In search of a fundamental theory.” *Oxford Paperbacks*, 1997.

Dennett, Daniel C. “Consciousness Explained.” *Little, Brown & Company,* 1991.

Dennett, Daniel C. “Elbow Room: The Varieties of Free Will Worth Wanting.” *MIT Press*, 1984.

Dennett, Daniel C. “Freedom Evolves.” *Viking Penguin*, 2003.

Dogen, Eihei. “The Heart of Dogen’s Shobogenzo.” Translated and annotated by Norman Waddell and Masao Abe. *State University of New York Press*, 2002.

Eliasmith, Chris, and Charles H. Anderson. “Neural engineering: Computation, representation, and dynamics in neurobiological systems.” *MIT press*, 2003.

Eliasmith, Chris, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. "A large-scale model of the functioning brain." *science 338, no. 6111*, 2012: 1202-1205.

Eliasmith, Chris. "Attractive and in-discrete." *Minds and Machines 11.3*, 2001: 417-426.

Eliasmith, Chris, Jan Gosmann, and Xuan Choo. "BioSpaun: A large-scale behaving brain model with complex neurons." *arXiv* preprint arXiv:1602.05220, 2016.

Eliasmith, Chris. “How to build a brain.” *TEDx,* 2013. https://www.youtube.com/watch?v=g2HHJfovb5E

Eliasmith, Chris. “How to build a brain: A neural architecture for biological cognition.” *OUP USA*, 2013.

Eliasmith, Chris. "Moving beyond metaphors: Understanding the mind for what it is." *The Journal of philosophy 100.10*, 2003: 493-520.

Eliasmith, Chris. "The myth of the Turing machine: The failings of functionalism and related theses." *Journal of Experimental & Theoretical Artificial Intelligence 14.1*, 2002: 1-8.

Ellis, George FR. "Physics in the real universe: Time and spacetime." *General relativity and gravitation 38*, 2006: 1797-1824.

Harris, Sam. “Free will.” *Simon and Schuster*, 2012.

Harris, Sam. “Waking up: A guide to spirituality without religion.” *Simon and Schuster*, 2014.

Hemmo, Meir, and Orly Shenker. "A challenge to the second law of thermodynamics from cognitive science and vice versa." *Synthese 199*, no. 1-2, 2021: 4897-4927.

Hossenfelder, Sabine. “The Problem of Now.” Backreaction, *Blogger,* 2014. http://backreaction.blogspot.com/2014/04/the-problem-of-now.html.

Libet, Benjamin. “Mind time: The temporal factor in consciousness.” *Harvard University Press*, 2004.

McKeon, Richard. ed. “The basic works of Aristotle.” *Modern Library*, 2001.

Mele, Alfred R. “Free: Why science hasn't disproved free will.” *Oxford University Press*, 2014.

Miller, Kristie. "Presentism, eternalism, and the growing block." *A Companion to the Philosophy of Time*, 2013: 345-364.

Nishitani, Keiji. “Religion and nothingness.” *University of California Press*, 1982.

Ramachandran, Vilayanur S. “The Emerging Mind.” The Reith Lectures, The Royal Institution of Great Britain, London England, 2003. Lecture.

Ramachandran, Vilayanur S. “The tell-tale brain: Unlocking the mystery of human nature.” *Random House Inc.*, 2012.

Savitt, Steven. "Being and Becoming in Modern Physics." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Winter 2021 Edition. https://plato.stanford.edu/archives/win2021/entries/spacetime-bebecome

Schultze-Kraft, Matthias, Daniel Birman, Marco Rusconi, Carsten Allefeld, Kai Görgen, Sven Dähne, Benjamin Blankertz, and John-Dylan Haynes. "The point of no return in vetoing self-initiated movements." *Proceedings of the national Academy of Sciences* 113, no. 4, 2016: 1080-1085.

Searle, John R. “Biological Naturalism.” In The Blackwell Companion to Consciousness. Edited by Max Velmans and Susan Schneider. *Blackwell Publishing Ltd*, 2007: 325-334.

Searle, John R. "Minds, brains, and programs." *Behavioral and brain sciences 3.3,* 1980: 417-424.

Spekkens, Robert. “The Quantum Physicist as Causal Detective.” *The Philosophy of Quantum Theory*, The University of Waterloo, Waterloo Canada, 2003. Lecture.

Turing, Alan. "Computing machinery and intelligence.” *Mind 59.236*, 1950: 433-460.

Van Gelder, Timothy, and Robert F. Port. "It’s about time: An overview of the dynamical approach to cognition." *Mind as motion: Explorations in the dynamics of cognition 1*, 1995: 43.