Cyborg Minds

Analyzing the necessary conditions for testing functionalism through an experimental philosophy of neurotechnology

Freek van der Weij

Master Thesis

July 26, 2022

Supervisors: dr. Y.J. Erden &
prof. dr. ir. P.P.C.C. Verbeek

University of Twente, Faculty of
Behavioural, Management, and
Social Sciences
Enschede, the Netherlands
MSc Philosophy of Science,
Technology and Society - PSTS

Table of contents

Summary

Philosophy of mind is typically an abstract, theoretical field of study which does not lead to empirical predictions. I argue that a theory of mind is stronger if it results in empirically testable hypotheses. This thesis specifically investigates the necessary conditions for testing functionalism. It does so by exploring how an experimental philosophy of neurotechnology could be used to this end, which entails that I analyze what is needed to test functionalism using neurotechnology.

Functionalism generally holds that what makes something a mental state solely depends on its function or the role it plays in the cognitive system of which it is a part, not on its internal constitution. However, functionalism is a general approach that encompasses multiple theories. The first chapter of this thesis explains three main functionalist theories, namely machine state functionalism, psychofunctionalism, and analytic functionalism. I decide to use psychofunctionalism in the rest of the argument. Psychofunctionalism is based on concepts from cognitive science and defines mental states according to how they relate to inputs, outputs, and other mental states. I consider psychofunctionalism to be the functionalist theory that is best able to deal with objections to functionalism, while also allowing for multiple realizability.

Multiple realizability means that mental states do not depend on particular physical characteristics, but instead can potentially be realized by biological as well as artificial agents. I discuss functionalism in the first place because it is commonly associated with multiple realizability, which, I argue, creates potential for empirical testability.

I hypothesize that psychofunctionalism should result in the hypothesis that someone's mental state will not be affected when a cognitive function is executed by a neurotechnology instead of a biological mechanism. This hypothesis could be tested in Schneider's Chip Test, in which such a scenario is created, and participants are asked to report on their experience regarding the substituted function. I propose a cognitive hippocampal prosthesis as an example of a neurotechnology to be used in this Chip Test.

Next, I raise potential problems that could arise when executing the Chip Test in this way. I explain how one can deal with criticism regarding reliance on phenomenal reports and how test results can be interpreted, among other topics. I furthermore determine that a cognitive hippocampal prosthesis, and presumably existing technology in general, is not sophisticated enough to use in the Chip Test.

In the final chapter, I formalize the necessary conditions for empirically testing (psycho)functionalism. These conditions are formulated partly as proposed improvements to psychofunctionalist theory. I argue that psychofunctionalism should specify the level of detail at which a function is to be substituted, and that it should endorse extended mind theory. Moreover, the final chapter describes the conditions required for using a neurotechnology in the Chip Test. I explain that a neurotechnology should fully replace a cognitive function that is general enough for someone to report on, while replicating that function at a high level of detail.

With this thesis, I aim not just to contribute to the philosophy of mind, but also to gather insights that can be used in analysis of the ethical dimensions of neurotechnologies. Particularly, if a neurotechnology acting as a replacement of a function can lead to distorted mental states, such a scenario may involve serious harm. This is something to be considered in the development and regulation of such technologies, especially in the light of the increasing number and variety of neurotechnological devices.

Acknowledgements

During the process of writing this thesis, I have been supported and motivated by a lot of great people, to whom I would like to express my gratitude.

First and foremost, I would like to thank my first supervisor, Y.J. Erden. Right from the start, I have found that her dedicated and structured supervision style suits me very well. However, she did not just feel like my thesis supervisor, but also like my mentor and career coach. Besides, by enabling me to present on a collaborative project at the Neurotechnology Meets Artificial Intelligence conference in Munich, she has directly contributed to my career beyond PSTS. I want to thank her for helping me make my interests explicit and giving me confidence to pursue them.

I would also like to thank my second supervisor Peter-Paul Verbeek, whose involvement in this project I have appreciated a lot as well. His broad knowledge has allowed me to better situate my thesis topic in relation to different fields of philosophy. Besides, he provides feedback in such a clear, constructive, and open-minded way that it is a joy to receive.

I am furthermore thankful for my fellow students, in particular Bouke, Lauren, and Maaike, for the interesting discussions and for making PSTS a fun experience, despite the impact of the pandemic.

Some other people who cannot be left out in this section are the rest of my friends, my parents, my brothers, and Anniek. Besides allowing me to improve the clarity of my thesis structure during conversations, they have provided the distractions that were necessary to keep me motivated throughout the thesis writing process.

Introduction

Philosophy of mind is a branch of philosophy that has existed for many centuries, at least going back to the mind-body dualism discussed in the fourth century B.C.E. by Plato (ca. 390-385 B.C.E./1997). From that period until the present, a great variety of philosophers have studied philosophy of mind, developing new theories and criticizing old ones. The mind has however always both literally and figuratively remained an ungraspable concept. Many discussions have centered around thought experiments like Descartes' (1641/2008) evil demon, the phenomenal zombie mentioned by Chalmers (1996), or the Chinese room argument made by Searle (1980). Valuable progress has been made by considering these hypothetical scenarios. Still, with this kind of method, philosophy of mind in many cases has remained an abstract, theoretical field of study. It has led to a range of hypotheses, but often, these have not been tested based on empirical data. In certain respects, philosophy of mind could potentially be expanded to become a more scientific research field, at least according to the Popperian view that science is distinguished from non-science through its ability to provide falsifiable hypotheses (Popper, 2014). This idea provides the foundation for this thesis.

To be more specific, this thesis investigates necessary conditions for the empirical testability of functionalism through an experimental philosophy of neurotechnology. These terms need to be explained. Functionalism is an approach to the mind that, I argue, is suited for this thesis since it grants a relatively clear relation to empirical claims. Functionalism holds that mental states depend on the execution of certain functions, not on particular physical characteristics (Levin, 2021). Thereby, functionalism generally allows for *multiple realizability*. Multiple realizability means that systems made of material that differs from the biological makeup of humans can, at least in principle, have similar mental states. As a consequence, functionalism hypothesizes that agents that are functionally equivalent to humans, will have similar experiences. Moreover, it entails, as I argue later in this thesis, that a person that uses inorganic technology to execute a function in the same way as it is normally executed by the brain, should have similar mental states as well. There is little literature that discusses this idea. The sources I found that do, are mentioned in chapter 3. Before recent years, such a partial replacement was probably likely considered to be a science fiction scenario. Now, research on neurotechnology that aims to replace or restore cognitive functions has grown significantly, leading to an increase in the number as well as the variety of available

neurotechnologies (Vázquez-Guardado et al., 2020). This makes partial replacement not only seem like a near-future possibility, but potentially even as an already existing procedure.

With this development, the impact of partial replacement on a person's mind becomes an increasingly urgent research subject as well. This research should thereby not only be seen in the light of philosophy of mind but also with regard to the ethical analysis of neurotechnologies. If the functionalist hypothesis regarding the effects of a neurotechnology on the mind turns out to be inaccurate, this can have significant and potentially harmful consequences for the people using these tools. In that case, partial replacement of cognitive functions might lead to distorted mental states that negatively impact people's well-being. This relationship between philosophy of mind and the ethics of neurotechnology has not yet been given sufficient attention in literature, I argue.

To gain insights regarding philosophy of mind and the ethical consequences of neurotechnologies, I propose to study the empirical testability of functionalism through an 'experimental philosophy of neurotechnology'. Haselager et al. plead for this method in their recent paper (Haselager et al., 2021). They particularly explain how neurotechnology could help to understand the factors influencing one's sense of agency. However, they also argue more generally for using neurotechnology in experimental philosophy. Experimental philosophy integrates questions that are traditionally associated with philosophy with common approaches from psychology and cognitive science (Knobe & Nichols, 2017). In this way, experimental philosophy is used to provide empirical evidence in the study of philosophical questions. Haselager et al. use the following argument for an experimental philosophy of neurotechnology:

> *"Neurotechnology can enable us to create and/or imagine cases where traditional concepts reveal aspects of their usage 'under stress', so to speak, which could lead to a better understanding of when, how, and why such concepts apply, or to reveal cracks in our understanding that may also have consequences for their application under normal circumstances. Hence, neurotechnology may provide opportunities for what we would like to call 'empirically guided thought experiments'; extrapolations or imaginations of (near-)possible conditions that, if well chosen, could illuminate our thinking about mind, ethics, law, etc." (Haselager et al., 2021, p. 61)*

So, they argue that we can improve our understanding of philosophical concepts by analyzing special cases that can be realized in practice with neurotechnology. I hold that an experimental

philosophy of neurotechnology will be useful to examine the necessary conditions for empirically testing functionalism, which is why this method will be central to this thesis.

First, in the following section, I describe what I mean with terms like consciousness, mind, and mental state. Then, in chapter 1, I go into detail on different functionalist theories of mind, namely machine state functionalism, psychofunctionalism, and analytic functionalism. Chapter 2 covers objections to these theories and selects a version of functionalism that is as strong as possible while at the same time allowing for empirical testability. Chapter 3 describes the Chip Test proposed by Susan Schneider as a method for testing functionalism. This method relies on phenomenal reports of people whose cognitive functions are (partially) performed by inorganic technology. In that chapter, I also illustrate how a cognitive hippocampal prosthesis could function as a concrete example of how the Chip Test might be executed using state-of-the-art technology. This technology is being developed to restore memory functioning for people who have lost certain memory capacities. Chapter 4 addresses objections that could be raised against the proposed testing method and analyzes the suitability of using a cognitive hippocampal prosthesis in this test. I argue that both functionalist theory as well as a neurotechnology like the cognitive hippocampal prosthesis are currently insufficiently developed for an empirical test of functionalism based on the Chip Test. Finally, the last chapter, chapter 5, explains how functionalist theory as well as neurotechnology should change to fulfill the necessary conditions for the empirical testability of functionalism. These proposed changes are derived from the analysis of potential issues with the Chip Test in chapter 4.

Consciousness, mind, or mental state?

It is important to start by making clear what I mean when using terms like mind, mental state, and consciousness, since there is little consensus on what these terms mean and how they relate to each other, at least in a philosophical context. In this section, I will provide a general description of how these terms are used in this thesis, while I stress at the same time that a formal, uncontroversial definition is impossible. In the final section of this chapter, I specifically consider recent theories which argue that the mind does not only depend on the brain.

The words ‘conscious’ and ‘consciousness’ are overarching terms used to describe a wide range of mental phenomena. Consciousness can be meant in a political or social sense (i.e., being aware of certain social issues), or relate to understanding, but I focus on consciousness as a philosophical concept. Regarding its meaning in that sense, the *Stanford Encyclopedia of Philosophy* entry on consciousness does not even provide a general definition,

and merely mentions that consciousness "lacks any agreed upon theory" and that "no aspect of mind […] is more puzzling" (Van Gulick, 2021). Still, the entry provides a useful way to categorize distinct meanings ascribed to the term consciousness. According to the author of this entry, a difference can be noticed between 'creature consciousness' and 'state consciousness' (Van Gulick, 2021). Creature consciousness refers to a characteristic of a system as a whole. So, the question of creature consciousness is about whether an animal, human, or AI system is conscious. State consciousness on the other hand refers to being aware of a mental state one is in. Questions about state consciousness thus concern features of consciousness, instead of whether certain agents have consciousness at all.

I will go over some definitions provided in other works to illustrate this categorization. According to the *Internet Encyclopedia of Philosophy* (Gennaro, n.d.), the most commonly used definition describes consciousness as "what it is like" to be in a certain mental state, using Thomas Nagel's phrase (Nagel, 2013). This refers to a version of state consciousness, because it describes the subjective quality of being in a particular state, not the idea that one is a conscious agent in general. The *Merriam-Webster* (n.d.) dictionary shows more examples of definitions of consciousness, like "the quality or state of being aware especially of something within oneself", "the state or fact of being conscious of an external object, state, or fact", or "the state of being characterized by sensation, emotion, volition, and thought". The first two definitions in the *Merriam-Webster* dictionary I mentioned describe a form closest aligned with state consciousness, while the third concerns a general characteristic of an agent and therefore aligns more with creature consciousness.

Within the two categories of consciousness mentioned, many more subtypes can be discerned, and an agreed-upon definition remains out of sight. This leads to a potential lack of clarity when using the term consciousness in academic work. In the *Stanford Encyclopedia of Philosophy* entry on functionalism (Levin, 2021) for example, consciousness seems to be avoided for this reason. The works referenced in this entry often have a title that contains consciousness or are written in the journal of consciousness studies, but Levin only talks of 'mental states' and 'minds' as the subject matter of functionalism. He also defines functionalism as a theory of mental states, not as a theory of consciousness as some others do (e.g., Heil, 1998; Chalmers, 1996).

I argue that, for the aim of this thesis, totally avoiding the term consciousness would lead to an unnecessary yet significant loss in clarity. This is different for this thesis compared to a work like Levin's encyclopedia entry on functionalism (2021), since I refer to literature related to the Chip Test. In these works, consciousness of AI is a central topic, making it hard

to circumvent the term. I will however make as clear as possible how I use the term consciousness. When I mention the term consciousness by itself, it refers to creature consciousness, so the general feature of consciousness of an agent. I use this as a synonym of, and thus interchangeably with, the term mind. In its definition of consciousness as "the state of being characterized by sensation, emotion, volition, and thought", Merriam-Webster (n.d.) also considers mind as a synonym. So, when I say that a human has a mind, I mean that a human has creature consciousness.

When I write that an agent is conscious *of* a particular phenomenon or consciously experiences something, I refer to state consciousness. The term mental state is used in a similar way, since I hold mental states to refer to different elements of consciousness. For example, one can be in the mental state 'in pain', while at the same time being in the mental state 'hungry'. Thus, I use the term mental state as an element of experience that one can be conscious of.

According to Van Gulick (2021), "as long as one avoids confusion by being clear about one's meanings, there is great value in having a variety of concepts by which we can access and grasp consciousness in all its rich complexity". This is what I have tried to achieve in this section, by explicating the difference between creature and state consciousness as well as the words I use to refer to these concepts (i.e., I use consciousness or mind for creature consciousness, and argue that one can also be conscious *of* a particular mental state). This should be sufficient for my thesis, especially since it focuses on *how* consciousness arises ('The explanatory question' according to Van Gulick, 2021), instead of *what* consciousness is ('The descriptive question'). Still, these questions remain dependent on each other.

The mind beyond the brain

There is more to discuss regarding terms like 'consciousness', 'mind', and 'mental state' than their general meaning as described in the previous section. It is particularly important to consider different theories of what constitutes cognition, which I hold to be what the mind does[1]. Cognitivism is the theory that cognition can be understood as *internal* manipulations of representations (Adams & Aizawa, 2010). This view generally holds cognition to be confined

[1] In a similar way, it is widely assumed, according to Trigg & Kalish (2011), that cognitive science aims to understand how the mind works. I recognize that scholars like Trigg & Kalish discuss confusions on the relationship between the terms 'cognition' and 'mind' as well. However, to address these would go beyond the scope of this thesis. Such fundamental philosophical concepts are bound to lack straightforward definitions. By distinguishing the mind as a static feature from cognition as a process, I argue that these terms have sufficiently been made clear.

to the brain (or central nervous system). Thus, according to cognitivism, cognitive processes inside the brain typically interact with noncognitive biological, chemical, and physical processes external to the brain. Thinking is considered to happen independently of external factors and merely consists of the brain connecting sensory inputs to behavioral outputs (Newen et al., 2018, p. 5). As a consequence, one can theoretically have full understanding of cognition by only studying the brain. However, multiple scholars dispute this 'traditional' view on cognition (as it is referred to by Rowlands, 2010 & Menary, 2010) and argue that certain extracranial processes should instead also be considered to constitute cognition.

Four main theories exist regarding how cognitivism should be expanded. This field has been referred to as '4E cognition' (Newen et al., 2018). '4E' here refers to embodied, embedded, enacted, and extended cognition. These approaches describe different aspects that cognitivism is held to lack. Embodied cognition holds that cognition is constituted by bodily processes besides those in the brain (Rowlands, 2010). So, this approach argues that cognitive processes occur through interaction between the brain and the rest of the body. Enacted cognition is the idea that cognition also involves moving the body to act on one's environment. Embedded cognition holds that cognitive processes generally depend on the external environment to function the way they do. The environment enhances our cognitive capacities, according to this approach. Finally, extended cognition holds that environmental elements outside the body can be part of cognition. Extended cognition is similar to embedded cognition yet diverges more radically from cognitivism. It argues that cognition can not only *depend* on elements outside the body but also be *constituted* by them.

In the empirical test that I will propose, I intervene in human cognition by connecting the brain to a neurotechnology. However, according to theories like 4E cognition, the mind has always been dependent on, if not constituted by external technologies. In chapter 4 and 5, I will further discuss the implications of 4E for the aim of my thesis.

In these previous two sections, I have explained what I mean when using fundamental terms like mind, consciousness, and mental state in this thesis. To this end, I have also mentioned contemporary theories that argue for expanding cognition beyond the brain. The next chapter will discuss functionalism, the theory of mind that is central to this thesis.

Chapter 1: What is functionalism?

General features of functionalism

To be able to investigate the empirical testability of functionalism, I first need to explain how I understand the term. Functionalism encompasses a variety of theories of mind (Block, 2013), and in this first chapter, I provide an overview of them. I first briefly introduce some general characteristics of functionalism. I follow this up with an explanation of two important preceding theories of mind which functionalist theories are a response to, namely behaviorism and type identity theory. These will help to understand the problems that certain features of functionalism try to resolve. I show this first by describing the earliest version of functionalism, i.e. machine state functionalism, as put forward by Putnam (1975a; 1975b). This theory uses the Turing machine as a metaphor to equate mental states with the entire structural state of a system. This section is followed by an illustration of psychofunctionalism, which describes mental states using terms from cognitive psychology, and of analytic functionalism, which relies on 'common sense' psychological terms (Levin, 2021).

In philosophy of mind, functionalism is a general approach that encompasses multiple functionalist theories which emerged in the second half of the 20th century along with improvements in the field of computation (Heil, 1998). The computer showed that systems built on inorganic mechanisms could perform an increasing number of operations that only humans were thought to be capable of before, like performing complex computations and communicating large amounts of information. Heil (1998) supposes that this partly inspired functionalism, since functionalists generally hold that what makes something a mental state depends "not on its internal constitution, but solely on its function, or the role it plays, in the cognitive system of which it is a part" (Levin, 2021). The computer science terms 'software' and 'hardware' became important metaphors for this theory: they show how certain functions, like computations, can be described without referring to their 'realization' in material systems. For example, we can ascribe the functional property of being able to calculate 20 x 13 to both (certain) humans as well as calculating machines. Functionalists generally argue that, like with computations, one can speak of mental states at this 'higher level' as well. On this account, the relation between mental states and the material world is similar to the relation between computer programs (the software) and the device on which they are run (the hardware).

A mental state, according to functionalist theories, is defined by how it is causally related to inputs (sensory stimulations), outputs (behavior), and other mental states. This might

seem to lead to circularity, since the definition of the term includes the term itself. Nevertheless, functionalism aims to resolve this circularity by defining every mental state at once, giving a complete explanation of mental states and how they each relate to environmental factors and behavior (Heil, 1998, pp. 100–102).

Thus, functionalism is a general approach that spans multiple different projects (Block, 2013). Therefore, when I use the term functionalism, one should recognize that it encompasses a variety of functional theories of mind. In the next section, I will start with a short description of theories of mind which functionalist theories aim to improve upon.

Antecedents of functionalism

It is helpful to first discuss antecedents of functionalism to make clear what issues functionalism intends to resolve and how well it succeeds in doing so. One important antecedent to functionalism is behaviorism, which originated in the middle of the 20th century (Polger & Shapiro, 2016). Behaviorist theories include the empirical psychological theories linked mainly to Watson and Skinner as well as the later 'logical' or 'analytical' behaviorism of philosophers like Malcolm and Ryle, according to Graham (2019). The philosophical theory of logical behaviorism is most relevant for this chapter, since it is a theory of what mental states are, just like the functionalist theories discussed in the following sections. Logical behaviorism holds that mental states can be fully explained through statements about behavioral dispositions (Graham, 2019). I could for example propose 'person X moans after having eaten spoiled food' as a (simplified) disposition for stomachache. An important argument put forward against using dispositions to explain the mind, is that many examples can be shown whereby someone does not act on a disposition, while being in the related mental state and the other way around. For example, someone can experience pain, but suppress the behavioral dispositions of pain, like shouting or crying. For a full picture of a mental state, reference to other mental states is argued to be necessary and this was therefore included later in functionalist theories (Levin, 2021).

A second relevant antecedent to functionalism is called type identity theory. It was developed (Smart, 1959) and consequently criticized (Putnam, 1975a, 1975b) before functionalist theories, according to Polger & Shapiro (2016), became popular at the end of the 20th century. Type identity theory holds that mental states can be grouped into types which can be equated to types of brain states. In this way, type identity theory connects mental states to the biology of the brain, disallowing agents with a different physical makeup to have similar mental states (Smart, 1959). It is a reductionist approach: the mental is reduced to the physical. Functionalism aims to oppose this reductionism, by allowing for *multiple realizability*, meaning

that similar mental states can be physically realized in different ways (Heil, 1998, p. 99). This idea was explained first by Putnam (2013). In the next section, I will explain further how Putnam's machine state functionalism aimed to overcome the limitations of type identity theory.

Machine state functionalism

Machine state functionalism (also referred to as 'computational functionalism') is the functionalist theory Putnam advanced in the earlier phase of his academic career (particularly explained in Putnam 1975a; 1975b). It was built on the concept of a Turing machine: an abstract system, which 'reads' input from a 'cell' of a 'tape' and changes that cell into a symbol (the output) based on the instructions in a 'machine table' (Putnam, 1975a). It moves to the next cell of the tape, and the cycle starts again. These elements of the Turing machine are in quote marks because the terms are used figuratively: a Turing machine can be configured in varying ways. To understand this, the distinction between the *logical state* of the machine and its *structural state* plays a crucial role. The description of a logical state does not specify the physical characteristics of the system, while the structural state of a system does. The logical state merely concerns an abstracted representation of the state of a system which can theoretically be realized using a range of materials (for example either organic or artificial). Different structural states could as a result be described with the same logical state.

Putnam argues that we can compare logical states of Turing machines to mental states of humans. He states that they are both descriptions of the functional organization of systems at a 'higher', abstract level that does not require the specification of physical characteristics, thereby allowing for multiple realizability (Putnam, 1975b). Thus, he uses the metaphor of a Turing machine to argue that mental states are not brain states, but instead a functional state of a whole organism. This functional organization of a system then describes how inputs, outputs, and the possible functional states of a system relate to each other. Using this approach, Putnam especially aims to overcome the consequence of identity theory that no agents other than humans can have similar mental states (Shagrir, 2005). This is the multiple realizability feature that initially attracted Putnam, myself, and other scholars to functionalism (Jaworski, 2008; Putnam, 2013). Besides, it is an effort to resolve the issue with earlier behaviorist theories which could only reduce mental states to relationships between inputs and outputs and not through reference to another mental state.

A major objection to machine state functionalism in particular is that logical states are total states of a system. Therefore, there is no room provided for multiple distinct internal states

that are realized simultaneously. There is no room to both be in the mental state ‘hungry’ and the state ‘happy’ for example, in contrast to our daily experience. This is the downside of only machine state functionalism defining the state of a system as a whole. This issue in particular led to decreased importance of and attention for machine state functionalism, according to Levin (2021).

Psychofunctionalism

Where machine state functionalism derived its methods mainly from developments in computation, psychofunctionalism, originally associated primarily with Fodor, stems primarily from cognitive psychological theories (Levin, 2021). Psychofunctionalism describes mental terms on the basis of how they relate to inputs, outputs, and other mental terms, according to Block (2013, pp. 268–270). These descriptions are conjoined into a so-called Ramsey sentence. To provide a highly simplified example, someone being hungry means that if there is a lack of food coming into the body, this could lead to a walk to the refrigerator, if there is not a mental state like wanting to diet that discourages the behavior. To construct a full Ramsey sentence, the mental state of ‘wanting to diet’ would consequently also have to be defined with references to inputs, outputs, and other mental states, and so forth until all mental states have been defined and connected in this manner.

What characterizes psychofunctionalism in this practice is that the mental states that are referred to, are concepts from cognitive psychology (Block, 2013, p. 269). They do not just derive from any given explanation, but instead from the best *scientific* explanation of human behavior. Terms used in this explanation can be inspired by folk psychology, meaning the psychological explanations given by humans in daily life, also referred to as ‘common sense’. However, experiments in cognitive psychology can lead to a theory that uses terms which diverge from the folk psychological terms which analytic functionalism relies on for describing mental states (Levin, 2021). One of the historical examples provided by Churchland (2013, p. 44) is that people with psychotic episodes were centuries ago often seen as being possessed by demons, or as witches. These folk psychological concepts have gradually been replaced by new terms grounded in cognitive science research.

It could thus be that progress in cognitive science will make psychological concepts scientifically irrelevant even though they are currently used in a way central to daily life. It can be noted that although some philosophers think this distinction between cognitive psychology and folk psychological terms will diverge drastically (Churchland, 1981), Levin (2021)

estimates that most scholars think they will remain relatively comparable. Nonetheless, the issue remains disputed.

Psychofunctionalism, just like machine state functionalism, allows for a definition of a mental state in reference to the current state of an agent. For that reason, it has an advantage over behaviorism, which reduces mental states to behavioral dispositions that only consist of input-output relationships. Psychofunctionalism is in contrast with machine state functionalism in the sense that psychofunctionalism provides a method for arguing that an agent is in multiple mental states at once. As a consequence, it provides a way to describe interactions between mental states, like how being nervous suppresses the urge to eat. For that reason, psychofunctionalism overcomes a major objection to machine state functionalism, which considers the machine table as one general state of the entire system and therefore cannot account for the existence of multiple simultaneously realized states.

Psychofunctionalism allows for multiple realizability as well, contrary to identity theory. The theory leaves room for different agents to have similar mental states, as long as they are able to perform certain functions as they are defined by cognitive psychology and related to certain inputs and outputs. As a consequence of this method, psychofunctionalism can define inputs and outputs not only as externally physical features observable and discernable by humans, but also for example neuronal inputs and outputs as they are discovered through scientific research (Block, 2013, p. 269). We cannot observe neurons directly through our senses. Yet, according to psychofunctionalism, they can be considered as elements of the definition of a mental state because scientific experiments can prove their existence. It might be that a mental state is only multiply realizable when it is executed in a way that is highly specific. If so, it could be that psychofunctionalism's tools for describing inputs and outputs are required for sufficiently detailed definitions of functions. In its dedication to scientific concepts, psychofunctionalism differs from analytic functionalism, the theory that is discussed next.

Analytic functionalism

Analytic functionalism, in contrast to psychofunctionalism, aims to restrict generalizations regarding mental states to an *a priori* analysis of concepts that align more closely with folk psychology (Levin, 2021). It aims to study the concepts people generally use to describe their mental states, without making use of cognitive psychology experiments. Analytic functionalism, mainly associated with authors like Lewis and Armstrong (Block, 2013, p. 296), holds that the meanings of mental state terms result from our non-scientific theories about them. Even though scientific methods might not be able to justify the use of these terms, that is not

considered necessary for the study of mental states. The rest of the method of analytic functionalism is comparable to psychofunctionalism and concerns the construction of a Ramsey sentence as well (Levin, 2021). So, again mental terms are defined by specifying their relations to inputs, outputs, and other mental terms. All definitions of mental terms are conjoined to construct a Ramsey sentence. Analytic functionalism in this way also overcomes issues with behaviorist theories by defining mental terms in relation to other mental terms. The difference between psychofunctionalism and analytic functionalism is that analytic functionalism uses folk psychological concepts as mental terms. Their reliance on folk psychology restricts analytic functionalists to rely on more limited means of specifying inputs and outputs in a Ramsey sentence (Block, 2013, p. 269). Where psychofunctionalism can make use of unobservable, scientific concepts like neuronal activity, analytic functionalism has to rely on externally observable physical characteristics for inputs and outputs. Furthermore, analytic functionalism is limited to mental terms which people can be *conscious* of, since only these are part of folk psychology. People cannot be conscious of specific neurons firing and therefore also not refer to them in a regular conversation.

Especially the earlier versions of analytic functionalism were a type of realizer functionalism (Heil, 1998, p. 98). To understand what that means, one must recognize the distinction between role and realizer functionalism. These are both classifications of functionalist theories that crosscut the classifications of theories so far. The distinction between role and realizer functionalism is based on what a theory holds the property of a mental state to be (Levin, 2021). Role functionalism concerns a mental state like pain to be the functional role pain plays in a larger system. Realizer functionalism on the other hand holds pain to be equated to how it is realized physically on a 'lower level', for example potentially in the stimulation of C-fibers. Earlier analytic functionalist theories strived for a theory in which common sense mental term types could be equated to types of physical states. However, this would lead to a functionalist theory that lacked the multiple realizability feature, making the theory vulnerable again to Putnam's critiques of identity theory (Putnam, 1975a, 1975b). The role versus realizer functionalism is an important distinction to consider in the next chapter, where I explain which functionalist theory should be considered in my investigation of the empirical testability of functionalism.

Chapter 2: Selecting a functionalist theory

Objections to functionalism

I have already raised some counterarguments to specific functionalist theories, but in this section, I will provide an overview of counterarguments to functionalism as a general approach. The second part of this chapter then evaluates what kind of functionalist theory is best able to deal with the mentioned objections from this chapter as well as the previous one. I thereby only consider functionalist theories that allow for multiple realizability, since a functionalist theory that excludes the possibility for multiple realizability cannot be tested using the experimental philosophy method that I propose later. That method relies on the possibility for mental states to arise as a result of executing a function using nonbiological means. I conclude in this chapter that psychofunctionalism is the strongest functionalist theory that also allows for multiple realizability.

One important objection to functionalism in general is that it is unable to sufficiently account for subjective qualitative experiences, also called *qualia*. The term 'qualia' describes what it is like to be in a certain mental state and for example experience emotions, perceptions, or bodily sensations. Authors have raised multiple counterexamples aimed to show how a purely functional description of a system does not provide a full account of qualia.

For example, Nagel (2013) famously argued that even if we have full knowledge of how an agent functions, we still do not know what it is like to be that agent. He gave the example of a bat and our lack of knowledge on how for example 'seeing' through echolocation is experienced. The *inverted qualia* argument describes a more specific example of what functional descriptions could lack. It holds that there could be two people who have the same functional descriptions, while nonetheless experiencing something different, e.g. experiencing as red what the other person experiences as green (Block, 2013, pp. 304–305). Another objection to functionalism is called the *absent qualia* argument. Chalmers (1996) advances this argument by considering 'zombies': agents that are functionally equivalent to humans but lack qualitative experience. Chalmers argued that such agents are conceivable, and that there is therefore an 'explanatory gap' between functional descriptions of a system and how they are experienced.

Another instantiation of the absent qualia objection is the so-called *China brain* argument first described by Ned Block (2013). Block sets up a thought experiment: what if the entire population of China (the country with a population size closest to the number of neurons

in a human brain) was made to act in a way that is functionally equivalent to a human brain? Every human would get a radio transmitter and receiver to behave as a neuron. If the transmitted radio signals, in an abstract sense, function in the same way as a brain's neural signals, this system could have 'software' identical to a brain. China would in this case be functionally equivalent to a brain, while also using humans as 'hardware', According to functionalism then, this system should be considered as having a mind. However, Block's argument goes, this is against our intuition of what a mind is. We would not ascribe a mind to a country as a whole, while functionalism would seem to lead to this conclusion. Block argues that the China brain argument shows that functionalism risks to be too liberal in what it considers to be a mind: it ascribes a mind to things that do not have them, namely to systems which contain parts that are conscious on their own.

A different counterargument to the functionalist approach to qualia is the *Chinese room* thought experiment by John Searle (1980). He imagined a closed room in which someone receives instructions which allow them to respond to Chinese characters as input with Chinese characters as output. This person outputs Chinese characters in a way indistinguishable from a native Chinese speaker through relying on these instructions. Someone outside the room would not be able to recognize that the person inside the room did not understand Chinese, even though the person in the room had not learned anything about the Chinese language before. Searle argues that the person who relies on the instructions cannot be said to *understand* Chinese, even though their inputs and outputs are functionally described in the same way as for a native Chinese speaker. Thus, for Searle, functional descriptions of a system are insufficient for the complete definition of understanding as a feature of the mind.

Dealing with objections

Extensive discussion is held in philosophical literature regarding these and other objections to functionalism. This chapter does not allow me to go into detail on all arguments and counterarguments, but I will establish a theory of functionalism that in my view optimally and sufficiently deals with the aforementioned major objections to functionalism.

One proposed way to respond to Searle's Chinese room argument and Block's China brain argument is a type of functionalism that has been developed by Daniel Dennett in particular. He argues for *homuncular functionalism*, which is a way of 'biting the bullet' that systems consisting of agents with mental states, can also have mental states themselves (Dennett, 2018). Homuncular functionalism holds that China in Block's thought experiment

can be said to have mental states, as well as that understanding can be attributed to Searle's Chinese room (taken as the whole system consisting of the room and the person in it).

However, this still does not provide a full account of qualia. One can wonder if there is any functionalist theory that does so, but I agree with Levin (2021), Block (2013) and Jacoby (1989) that a scientific, psychofunctionalist approach has the highest likelihood in succeeding here. Machine state functionalism in general faces objections mentioned before that have led the theory to fade in importance, like its lack of ability to define mental states in reference to other simultaneous mental states (Levin, 2021). Analytic functionalism lacks ways to differentiate between nuances in qualitative experiences because it only uses folk psychological terms to define mental states (Block, 2013, pp. 295–300). Besides, multiple historical examples can be found of ontologies based on folk psychology that have been disproven through scientific methods (Churchland, 2013, pp. 43-44). Psychofunctionalism instead uses scientific methods from cognitive psychology to find the best explanation of a mental state. Still, this method is limited to referring to information processing for explaining qualitative experiences.

I concede that qualitative experience thus remains an issue, also for psychofunctionalism. However, I am not convinced by the strength of 'conceivability arguments', as Jacoby (1989) calls them. These include objections to functionalism like the zombie or inverted qualia argument mentioned earlier. I am instead sympathetic with the argument made by Jacoby (1989) and others (Block & Stalnaker, 1999; Yablo, 2000) who argue that such conceivable scenarios do not entail actual possibility. That we can *think* of a zombie or inverted qualia, does not mean that they are *possible* scenarios that functionalism needs to account for. These types of hypothetical examples do not provide strong evidence of an explanatory gap in functionalism's definition of qualia[2].

Yet, an issue for psychofunctionalism remains regarding how 'chauvinism' can be overcome, as Block (2013) calls it. With that term, Block means withholding ascription of mental states from agents that have them. Since psychofunctionalism uses cognitive psychology to study humans and how human mental states can be defined as functions, this might lead to a limited view on what systems can have similar mental states. In this way, psychofunctionalism would lack the multiple realizability feature I look for in this chapter and become a variation of identity theory. For example, consider Martians that would be similar to humans when

[2] Being sympathetic with this line of reasoning does not mean that I consider all conceivability arguments to be useless. I merely aim to argue that conceivable scenarios are weaker in their argumentative strength than possible scenarios. As mentioned at the end of this chapter, it is neither possible nor necessary for the aim of this thesis to fully refute all objections to functionalism.

described in folk psychological terms of analytic functionalism. These might not have similar mental states to humans according to psychofunctionalism, because the exact scientific concepts cannot be applied to them. For example, they might have something that can be called 'memory', but they do not share the same distinction between short- and long-term memory. According to Block (2013), this seems counterintuitive. However, this would only be an issue when cognitive psychology research would lead psychofunctionalists to define mental states in terms that diverge strongly from folk psychological terms. This can only be assessed when science has developed further, so I do not consider this counterargument a strong enough objection against psychofunctionalism at this stage.

To uphold the multiple realizability feature of a psychofunctionalist theory of mind, it is important to consider how inputs and outputs of a function are defined. This can be done in functionalism using 'short-arm' description of inputs and outputs which contain reference to the sensory and motor system of an agent, respectively (Levin, 2021). This would mean that a function's input and output are defined based on specific biological features, severely limiting if not eliminating the possibility for multiple realizability. This issue can be prevented with a 'long-arm' description of inputs and outputs, where the input is defined in terms of external events and the output in terms of the agent's behavior. This loosens the conditions for the physical realization of functions, and thus is a preferred feature for a functionalist theory that aims to allow for multiple realizability[3].

I argue that a psychofunctionalist theory as described, which also recognizes the possibility of attributing mental states to systems containing minds, is most suitable for the rest of the thesis. I consider a role functionalism version of psychofunctionalism, whereby functions are equated to their role in a system, instead of their physical realizers. Furthermore, the psychofunctionalist theory should use 'long-arm' descriptions of the inputs and outputs of functions which do not put strict limitations on how a function is physically realized. In that way, functions might still be multiply realized. Besides, psychofunctionalism is regarded by proponents as a scientific approach to the mind (Block, 2013, p. 268). This means that its aims align best with the general aim of this thesis, which is to strive for a scientific functionalist theory by investigating empirical testability.

[3] Moreover, it accounts for the Twin Earth counterargument to functionalism raised by Putnam (1974). This thought experiment, according to Putnam, showed that the content of a mental state is insufficiently described without reference to the material world. He imagined a different planet ('Twin Earth') where a liquid existed that seemed to have the same characteristics as water has on Earth, yet it had a chemical formula that differed from H_2O. In this situation, Putnam argued, functionalism needs to mention the chemical formula of the liquid to be able to distinguish between the different contents of mental states of those on Earth and Twin Earth.

Questions understandably remain and much more can be written on objections and refutations of functionalism in general or specific functionalist theories, however I think this would not sufficiently suit the aim of my thesis. I do not intend to convince readers of functionalism in general or psychofunctionalism in particular. Instead, these first two chapters are only meant to establish the features of a version of functionalism that I consider to be strongest, and which can consequently be used in an investigation of the empirical testability of functionalism. The next chapter describes a proposal for how the psychofunctionalist theory I established in this chapter could be tested using current neurotechnology.

Chapter 3: Constructing an empirical test of functionalism

The Chip Test as an empirical test of functionalism

In the previous chapter, I set out a version of psychofunctionalism (which I hereafter refer to as PF) that I consider most suitable for an empirical test. In this chapter, I investigate what an empirical test of PF using current neurotechnologies could look like. First, I set out a tentative method for testing PF and its multiple realizability feature. This method is inspired by Susan Schneider's Chip Test in particular. The proposed procedure would involve a neurotechnology that is able to act as a substitution of a cognitive function of the brain. In the subsequent section, I first provide a quick description of what neurotechnologies are and which types of neurotechnologies currently exist. Then, I investigate which neurotechnology could potentially fulfill the necessary conditions for the test and thus be relevant for the aim of this thesis. I argue that a cognitive hippocampal prosthesis initially seems to be a good candidate.

A few works can be found in literature which discuss comparable tests. I will explain these and investigate their usefulness for this thesis. To start off, David Chalmers has written about an experiment that involves gradual replacement (Chalmers, 2016). Chalmers sketches a scenario where, at first, a small number of a biological brain's neurons are substituted by a functionally equivalent nonbiological system. Every neuron in the biological brain has a counterpart in the nonbiological system. This counterpart replicates the 'input-output behavior' of the neuron (Chalmers, 2016, p. 45), meaning that it returns the same output in response to a given input. This replacement process continues until all neuronal activity is substituted by inorganic activity. This experiment is supposed to be a test of functionalism, making Chalmers' paper useful to my thesis, although he seems to use the test more as a thought experiment and does not explicitly investigate empirical feasibility.

Similarly, Block (2002) explores how a 'superficial functional isomorph' could provide empirical contributions to knowledge on consciousness. This is a technological system "that is functionally isomorphic to us with respect to those causal relations among mental states, inputs, and outputs that are specified by 'folk psychology'" (Block, 2002, p. 399). Based on this quote, it seems that Block investigates empirical evidence for analytic functionalism in particular. One can tell since he describes functions in terms that are in accordance with folk psychology, not with terms from cognitive science as in PF. His paper focuses furthermore on testing consciousness in totally artificial systems, so that would only be the final stage of Chalmers'

gradual replacement experiment. Still, his work provides criticisms which are relevant to my proposed method for testing PF, and which will be addressed later in this chapter.

The only work I was able to find in literature that provided a useful description of an empirical test for functionalism is by Susan Schneider (2019). She has developed a 'Chip Test' for testing the possibility of AI consciousness. In this test, the brain activity in a certain physiological or functional region is suppressed, while a chip is connected to the brain at the same time. This chip is made to perform the same neurological behavior as used to be performed by the suppressed brain activity. Schneider then emphasizes the usage of introspection: the test should rely on people's reports about what they experience. One the one hand, if subjects are still able to provide introspective reports regarding the relevant function or physiological region, the chip is able to give rise to mental states, according to the Chip Test. On the other hand, they should also be able to notice and report to others if they lost consciousness of a mental state after replacement of a segment of the brain related to that mental state (Susan & Mandik, 2018). This person should notice such a 'substitution failure' in a similar way to people losing sight after an injury. The Chip Test can be repeated until the biological brain is fully substituted by chips. However, since I want to study whether such a test could be executed in the near future, I only study partial replacement, which presumably requires considerably less sophisticated technology. Moreover, fully or largely replacing a biological brain with nonbiological technology can be assumed to raise significantly more ethical issues.

In the rest of the thesis, I will use the Chip Test as the proposed method for testing PF. The test focuses on partial replacement, as I do, and provides a comprehensive description for collecting and interpreting information from the tested subject. The Chip Test is thereby a useful method to be investigated further in the rest of this thesis. To create a clearer picture of what such a method might look like, I will illustrate a proposal for a neurotechnology that initially seems to be suited to use as part of the Chip Test.

An example of a neurotechnology to use in the Chip Test

Before considering specific examples of neurotechnologies to serve in the Chip Test, I will explain more generally what neurotechnologies are and which types of neurotechnologies exist. Neurotechnologies are tools that are able to extract information from or feed information into the human nervous system, particularly the brain (Roelfsema et al., 2018). Some neurotechnologies both 'read' and 'write' neural information. By doing so, many neurotechnologies currently aim to restore cognitive functioning that was lost due to illness, disability, or injury. However, technological improvements might be used to cognitively

augment healthy people (Cinel et al., 2019). Neurotechnologies can be more or less invasive, which is dependent on whether they require introducing instruments into the body (Cinel et al., 2019, p. 2). Invasive neurotechnologies are generally characterized by the insertion of electrodes in the brain or on its surface.

An increasing range of technologies is becoming available for extracting and influencing neural information. Examples of neurotechnologies that read neural information are for example functional magnetic resonance imaging (fMRI) scanners, which create images of brain activity by measuring changes in blood flow, or electroencephalography (EEG), which concerns placing electrodes on the scalp to record electrical activity in the brain (Cinel et al., 2019, p. 2). However, such non-invasive neurotechnologies are not relevant towards the aim of my thesis of investigating the testability of PF. For that, I argue, a technology needs to be able to fulfill a function normally executed by the brain. Therefore, it needs to be able to receive inputs and convert those into outgoing signals. Such a technology should thus be able to both read and write information.

For that reason, I focus on neural prostheses, which are "assistive devices" that usually "restore functions lost as a result of neural damage" (Prochazka et al., 2001). They achieve this with neural signals as input, which is translated into electrical stimulation of particular nerves as output. This happens either through electrodes attached externally to the skin or through a device implanted inside the skull.

I thereby specifically look for a neural prosthesis that is able to fulfill a cognitive function as it would be described by PF. This means that it needs to act according to the scientific (not commonsense) understanding of that function. There are a number of neural prostheses that seem interesting in this regard which involve real-time speech synthesis for paralyzed individuals through connecting electrodes to the brain (Brumberg et al., 2010). These individuals are unable to translate certain neural signals into the motor action necessary for speech. The neural prosthesis is connected to a speech-related region of the brain, either by putting electrodes on top of the skull or implanting them. The detected signals are analyzed to infer what the individual would want to say, and those words are then expressed by a real-time speech synthesizer.

Despite replicating how speech functions in a healthy person, such a neural prosthesis would not be suited for the partial replacement test of PF I investigate in this thesis. I say this because it performs a function that only takes neural signals as input, while having speech created by the synthesizer as its output. It is therefore not fully integrated into cognition. It does read and write information, but it does not write *neural* information. Since the output is speech

instead of neural signals returning to the brain, relying on reporting would not make sense here. One's mental experience of the speech process will be limited, because such a neural prosthesis does not allow for neural feedback A person with this neural prosthesis would not be able to report on their consciousness of speech as a cognitive function, but only on the perception of hearing the words that are expressed.

An example of a neural prosthesis that is integrated into the cognitive system with both its in- and outputs, is a cognitive hippocampal prosthesis. This technology is also referred to by Schneider (2019) as an example of a chip that might be suitable for the Chip Test. A cognitive hippocampal prosthesis is a "cognitive prosthesis designed to restore the ability to form new long-term memories typically lost after damage to the hippocampus" (Berger et al., 2012, p. 198). The prosthesis is thus explicitly designed to perform a cognitive function that is normally executed by the brain. It predicts on the basis of neural input how to optimally output certain neural signals to form long-term memories. A hippocampal neural prosthetic has already been shown to function effectively for humans to restore short- and long-term memory encoding capabilities (Hampson et al., 2018). In overviews of the current state of neurotechnologies (Cinel et al., 2019; Prochazka et al., 2001; Vázquez-Guardado et al., 2020), I have not found an example of a neurotechnology that seemed more sophisticated and had neural inputs as well as outputs. The field is new and developing at a high pace, so a better alternative for the Chip Test is hard to establish but could currently exist. However, the cognitive hippocampal prosthesis mainly serves as a useful illustration for using modern technology in the Chip Test. It should therefore merely approximate the current state of technological development and I argue that it sufficiently does so.

In this chapter, I have sketched out the Chip Test as a potential empirical test for PF. Furthermore, I have given relevant examples of the present state of neurotechnology and explained why a cognitive hippocampal prosthesis could be a useful illustration of a modern technology that might be used as part of the Chip Test. The next chapter will describe potential objections to this way of executing the test, regarding the test's methodology as well as the selected technology.

Chapter 4: Potential problems for the Chip Test

Introduction

In this chapter, I analyze potential problems that could arise when executing the Chip Test using a cognitive hippocampal prosthesis in the way described in the previous chapter. In literature, some skepticism can be found that targets a Chip Test experiment like the one I consider (Udell & Schwitzgebel, 2021), and some other sources (Block, 2007; Dennett, 2001) that investigate consciousness of fully artificial agents will be useful in that regard as well. Still, overall, little has been written on an empirical test of functionalism in general or PF in particular that investigates partial replacement, so I will supplement existing literature with my own concerns regarding this method. This chapter first discusses concerns that could arise regarding the use of phenomenal reporting to gather data in the test. I argue that there is sufficient reason to trust verbal reports from subjects who have only a small part of their cognitive function replaced. Consequently, I treat potential problems related to how a function is defined in the Chip Test. I explain why I do not consider it an issue that subjects report in folk psychological terms, while PF describes functions according to cognitive psychology concepts. I furthermore argue how the Chip Test's outcomes can be interpreted as useful contributions to the debate on functionalism, PF in particular, even though it is unclear at what level of detail a function should be recreated by the technology. Then, I treat further potential problems regarding the interpretation of the test results, like the generalizability of results and the role of neuroplasticity. I end the chapter by reconsidering the current suitability of a cognitive hippocampal prosthesis for performing the Chip Test. I conclude that this prosthesis is insufficient for empirical testability of PF, since it does not fully substitute a person's memory function.

Issues for the Chip Test concerning phenomenal reporting

One crucial point of concern regarding the proposed Chip Test relates to how the researcher executing the experiment would gather empirical data. Schneider (2019) proposes that introspective reports could be used as evidence in the test. An introspective report would entail that the test subject tries to reflect on their experience of the function that is now executed by the chip. In the previous section, I have proposed that the 'chip' could currently be a neural prosthesis like a cognitive hippocampal prosthesis. Regarding such a technology, the test subject would be asked to reflect on how they experience their newly created memories as well

as memories that would not have been retained without the capacities enhanced by the prosthesis. For the second category, one can imagine checking beforehand how long the subject remembers a set of digits or words and whether this has changed.

A potential problem with such phenomenal reporting that Udell & Schwitzgebel (n.d.) mention is that someone who has their cognition replaced by technology cannot be trusted in their reports. They start by stating that those skeptical of the possibility of AI consciousness (and thus of the multiple realizability feature of PF) will not trust the reports of fully artificial agents. Schneider does not even trust those herself. Skeptics could argue that consciousness does not emerge from the mere execution of functions but instead requires a biological basis. Thus, they leave open the possibility for a functional isomorph of a human that provides the same reports without being conscious (this is similar to the phenomenal zombie described by Chalmers, 1996). Consequently, according to Udell & Schwitzgebel (n.d.), this view can also lead to doubt when only a part of the brain's function is replaced by artificial elements. These authors argue that the 'genuine' consciousness that we attribute to ourselves can be considered as a requirement for trustworthy introspective reports, especially by skeptics of multiple realizability. As soon as a human brain is not fully intact anymore, this could raise doubts on the genuineness of this person's consciousness. Therefore, the test would need to rely on consciousness as a precondition to convince these people, while that is exactly what the test tries to assess. This would undermine the Chip Test's method.

Nonetheless, I do not consider this to be a strong objection to the Chip Test. To explain why, I refer to the problem of other minds (Avramides, 2020). This philosophical problem revolves around the question "How can I know that the agents that I encounter and with which I interact, think and feel as I do?" This is relevant to the aforementioned objection about relying on phenomenal reports because that objection comes down to skepticism regarding the consciousness of other agents. One can experience their own thoughts and feelings, there seems to be no doubt about that. Descartes, with the famous expression "cogito, ergo sum", already equated one's essence to a 'thinking thing', the characteristic of the self that one can be most certain about (Descartes, 2008). However, can one also be so certain that others have a similar experience? Melnyk (1994) argues that one can reasonably hold that others have mental states on the basis of an inference to the best explanation. Other people not only show behavior that suggests that they experience thoughts and feelings, but they are biologically constituted in a largely identical way. Thus, because we know that we have certain mental states and are

biologically similar to others, we can reasonably infer from other people's recognizable behavior that others have similar mental states too[4].

Now, if one takes this account of justifying the attribution of mental states to other people, this provides an argument for weakening the objection that one cannot trust the reports of people with partial replacements. If only a function like memory is executed by inorganic means, the rest of the subject remains biologically very similar to a 'regular' human. Therefore, I argue that there is sufficient, or at least considerably stronger reason to trust that humans with partial replacement have similar mental states. As a result, there is also significantly better support for trusting phenomenal reports of these people in comparison to fully artificial agents.

Issues for the Chip Test regarding how a function is defined

The second main category of potential issues regarding the Chip Test relates to what qualifies as a function. Block (2007) would be critical of the Chip Test because of the difference between how functions are experienced and how they are defined by cognitive psychology. He mentions the following regarding a method like the Chip Test:

> *"[It] is in danger of focusing on the neural basis of higher-order thought […] rather than the neural basis of experiential content or even access to experiential content. To give an introspective report, the subject has to have a higher-order thought—so to insist on introspective reportability as the gold standard is to encourage leaving out cases in which subjects have experiences that are not adequately reflected in higher-order thoughts."* (Block, 2007, pp. 355–356)

Thus, Block notes that reporting on thoughts happens in folk psychological terms. Consequently, people are potentially unable to describe some parts of their experiences or at least not in high detail. Thereby, an incongruency might arise in the Chip Test I propose when a certain cognitive function is executed by a neural prosthesis. On the one hand, a function is defined according to potentially specific terms derived from cognitive psychology. On the other hand, those asked to report on their experience of this function are only able to report using general folk psychological terms. As a result, neurotechnology might replace a cognitive

[4] I recognize that much more can be said and has been said in other works on the problem of other minds. Avramides (2020) provides a comprehensive overview of literature on the issue, where multiple other solutions and related problems are listed. For the aim of this thesis, I argue that the solution I mention sufficiently deals with the main problem of other minds.

function at a scale that is too small for a person to recognize and report on. Humans are for example only conscious of the general experience of paying attention. However, attention also includes a pre-attentive stage where cognitive processes occur which people are not conscious of (Treisman et al., 1992).

Nonetheless, I think this incongruency can be resolved. For one, I estimate, in agreement with Levin (2021), that cognitive psychology terms will likely not diverge strongly from current folk psychological terms. However, I do predict on the basis of previous scientific development that cognitive science research will further distinguish specific parts of the larger functions we refer to in our commonsense reflection on cognition. In my view, cognitive science will construct a growing hierarchical functional ontology, where the most general function remains similar. For example, to stick with the function of memory: in daily discourse, we typically only use the general term 'memory'. Yet, cognitive science has been able to define different lower-level functions that make up what we in daily life refer to as 'memory', like short- and long-term memory and sensory processors. Some scientists posit the existence of intermediate-term memory as well (Kamiński, 2017). If we take the example of sight again, the same can be noticed. In folk psychology, we generally only use the word 'seeing' to describe the cognitive function of visual perception. Nevertheless, cognitive science has subdivided sight into lower-level functions like translating energy into neural signals, accommodation (changing the shape of the eye to focus on far or near objects), transmission of signals, etc. (Carlson, 2013).

Therefore, to resolve Block's skepticism regarding phenomenal reports, the Chip Test should include the replacement of a function at a level that people are able to report on. This means that if a commonsense function can be broken down based on cognitive science, it is insufficient if a technology only replaces one of these lower-level functions. All lower-level functions making up a more general function that is part of folk psychology should be executable by the technology to perform the Chip Test.

Even after having concluded this, questions remain about how a function should be defined. Again, according to PF, a function is defined on the basis of cognitive psychology concepts. However, this still leaves room for interpretation regarding the level of detail at which an inorganic substitution needs to be the same as its organic counterpart. Current literature shows disagreement on this topic. For example, Dennett (2001) argues that researchers who try to recreate the mind's functions underestimate the high level of detail that the definition of a function requires. He writes that he notices a bias towards 'functional minimalism' in science, according to which "less matters [in the execution of a function] than one might have thought"

(Dennett, 2001, p. 233). To exemplify this, Dennett refers to the example of a heart. A heart's function, according to functional minimalism, is merely to pump blood and everything that pumps blood could function as a heart, independent of the material it is made of. Thus, functional minimalism has loose conditions for two systems to be functionally equivalent.

Dennett (2001, p. 234), on the contrary, argues that functions depend on their 'micro-architecture', meaning the structure and mechanics at a micro-level. You cannot explain the mind, according to him, if you leave out mention of where elements relevant to a function are located and how they communicate at the neuronal level. Chalmers (2016), in his experiment on gradual replacement, investigates replacement of functions at this micro-level as well, although he does not explicitly mention that such fine-grained replication of a function is necessary for similar resulting mental states.

Schneider, in line with Chalmers, does not suggest having a particular view on, as she calls it, the 'granularity', meaning the level of functional detail, that is required for consciousness (Schneider, 2019, pp. 60–61). Instead, she proposes that the Chip Test can be used to discover the level at which functions are to be described for similar resulting mental states. A researcher should try out different levels of functional equivalence. The most general level of functional description that leads to similar mental states could be taken as the necessary granularity for PF. In this way, significant progress could potentially be made in the development of functionalist theory. However, Schneider writes that at a micro-scale, organic and inorganic systems will always be different, regardless of future technological development (Schneider & Mandik, 2018). She argues that the Chip Test is only suited for functional replacements at a lower level of detail.

Then, the question remains what a negative result, i.e., the total or partial loss of consciousness, would mean for PF. It can be that such a negative result shows an argument against PF, but this can also be explained away by arguing that the nonbiological replacement of a brain function does not function at the right scale. While replacements that are functional isomorphs at a more general level potentially already exist or will exist in the near future, functional isomorphs at a higher level of detail will take longer to be developed. When would a more detailed functional replacement be considered unachievable?

As a consequence, negative results could potentially be rejected for a very long time. A negative result on its own thus does not allow the experimenter to conclude much. Only after many unsuccessful tries at different levels of description would the Chip Test provide some evidence against PF. Still, this evidence would not be conclusive and could be turned around with new developments in cognitive science or the design of neurotechnology. Udell &

Schwitzgebel (n.d., p. 124) for this reason argue that Schneider's Chip Test "is a *sufficiency* test of consciousness rather than a necessity test: passing the test is held to be sufficient evidence for a confident attribution of consciousness to the target system, while failing the test does not guarantee a lack of consciousness". In the next chapter, I provide an attempt at limiting the level of detail required by PF for substitution of a function.

Finally, it is necessary to consider what exactly executes a function and what the consequences are of the discussion on 4E cognition mentioned in the introduction. It could be argued that the Chip Test unreasonably assumes that a function like memory is merely executed by the brain before a neural prosthesis is used. When this is assumed, the neural prosthesis would be the initiator of a connection between the brain and external factors. However, the 4E approaches would all reject the assumption that cognition is limited to the brain. These theories would argue that the body, action, or external elements are inextricably linked to the execution of a function (Newen et al., 2018). The Chip Test needs to recognize that without a neural prosthesis, there can already be reason to believe that it is not just a brain executing a function. Thus, the substitution of a function potentially requires not only the right relation to the brain, but also to the body and the external environment. Researchers executing the Chip Test would have to take these extracranial elements of a function into consideration when analyzing the functional similarity between a neural prosthesis and a brain function. Without this, a researcher might be convinced of functional similarity and as a consequence incorrectly interpret distorted experience of a mental state as an argument against PF.

Further issues concerning interpretation of results of the Chip Test

In the third section of this chapter, I discuss some further concerns regarding how empirical findings from the introspective reports are interpreted. I agree with Heil (1998, p. 16) that "questions that arise in the philosophy of mind are rarely susceptible to straightforward empirical investigation". An empirical finding does not self-evidently lead to a particular conclusion about a philosophical theory of mind. We should instead carefully reflect on this relationship. I have raised some potential problems already regarding the definition of functions, but some issues still remain to be discussed.

Firstly, I argue that PF should entail the hypothesis that a partial replacement of a brain's functions with an inorganic system would not lead to a total loss of or less complex consciousness. In line with Chalmers (2016, pp. 45-48), I consider three potential outcomes of the Chip Test. Firstly, consciousness might disappear completely when one cognitive

component is replaced. Chalmers deems it unlikely that a small component could have such a significant impact. Secondly, consciousness might gradually fade out when more components are replaced. Complexity would fade stepwise in this case. Chalmers considers it implausible as well that a system can exist which functions the exact same way as a biological system yet is not experienced similarly. Thirdly, consciousness could remain the same, even though the physical realization of cognitive processes of the system change. To Chalmers, this is the most likely scenario. He mentions that a partial replacement might be available much sooner, and this will, Chalmers argues, turn out to be convincing evidence already in favor of functionalism.

Scholars like Chalmers suppose that functionalist theories entail that partial replacement, if executed correctly, will not significantly affect one's mental states. Because PF allows for multiple realizability, the mental states related to the execution of a function should not depend on how the execution of this function is physically realized. However, it can be questioned whether current versions of functionalism like PF can indeed be held to entail this hypothesis. Although this seems to be a logical consequence, functionalism lacks mention of the 'cyborg': an agent that consists partially of biological, partially of nonbiological material. Therefore, multiple realizability of PF might not be empirically testable as the theory is defined now. This issue is discussed further in the next chapter.

Secondly, in a more general sense, the generalizability of results of the proposed Chip Test can be questioned as well, which is an issue that should receive significant attention. Generalizability is harmed when only few research subjects are investigated and when only one or a few technologies are tested as replacements. Therefore, I do not claim that a Chip Test like I propose with a cognitive hippocampal prosthesis would be a strong enough reason to prove or disprove PF. Still, I think that initial results can have some influence on the debate. This influence will grow if more neurotechnologies become suited for such a test and if a wider range of people get neural prostheses.

The third and final point that needs to be made here regarding the interpretation of the Chip Test's results concerns neuroplasticity. More specifically, one can question the effect of neuroplasticity when a function is not replicated exactly. Neuroplasticity refers to the ability for the nervous system, which the brain is part of, to change, for example through new neural pathways or changes in volume of brain regions related to certain functions (Costandi, 2016). If a technology performs a function similar to how the brain would, but not in exactly the same way, it might take some time for the remaining biological brain to adjust. The neuroplasticity of the brain could allow on the one hand for the brain to align better with the neural prosthesis

and maybe with time, consciousness of the function returns (Potter, 2010). This would potentially strengthen the Chip Test's evidence for multiple realizability.

On the other hand, neuroplasticity can potentially also have the effect that other parts of the brain that are not replaced, take over the function that the neural prosthesis is expected to play (Nordmann & Rip, 2009). Therefore, it might seem like the test subject slowly starts to experience the execution of the function of the neural prosthesis, while this function is in fact still being executed by the brain itself. This would weaken the Chip Test's evidence for multiple realizability. So, if the experience of mental states related to a supposedly replaced function returns over time, one cannot conclude whether this vouches for or against PF, since there are strong arguments for both positions.

Evaluating suitability of a cognitive hippocampal prosthesis for the Chip Test

Now that I have addressed potential issues with the methodology Chip Test, I will reflect briefly on the technology I initially proposed to use in that test, namely the cognitive hippocampal prosthesis. One can wonder whether this technology is sufficiently sophisticated to be a suitable candidate for such a test. On the one hand, a cognitive hippocampal prosthesis as described by Hampson et al. (2018) is built on a statistical model that aims to predict neural signals in the hippocampus related to successful memory functioning. The prosthesis then stimulates neurons in the hippocampus in a way that partially restores memory function for people with a medical condition like Alzheimer's disease. This way, the technology aims to function in a way that is similar to 'normal' memory at a level of description that seems sufficiently detailed.

On the other hand, the technology used in a Chip Test should fully substitute a function. The cognitive hippocampal prosthesis I refer to does function as an improvement to memory formation and retention generally, so both to short-term and long-memory. Thus, it operates at a level that patients can report on. Yet, the method prescribed in the Chip Test consists of fully shutting down a brain function and consequently substituting it with nonbiological technology. The cognitive hippocampal prosthesis does not take over the brain's memory functions, it merely supports and improves one's memory encoding and recall. This significantly complicates the interpretation of phenomenal reports in the Chip Test. If a large part of the memory function is still being executed biologically by the brain, what would one be able to conclude from people's reports on their experiences? If people report relatively similar experiences to the period before they had a neural prosthesis, this could also be ascribed to brain processes that were not yet damaged. Therefore, I hold that the cognitive hippocampal prosthesis described by Hampson et al. (2018) is insufficient to serve in the Chip Test.

In my research, I have not come across a neurotechnology which I considered to be more suitable for use in the Chip Test. Therefore, I argue that the Chip Test can currently not be executed. In the next chapter, I will more explicitly describe the features I deem necessary for a technology to serve in the Chip Test. But first, I will set out how the theory of PF should be adjusted to allow for empirical testability, based on the reflections described earlier in this chapter.

Chapter 5: Changes in theory and technology necessary for empirically testing functionalism

Creating an empirically testable psychofunctionalist theory

The final chapter of this thesis covers the implications of the previous chapter. The previous chapter discussed potential issues with a Chip Test of PF using a cognitive hippocampal prosthesis. I have also provided some ways in which these issues could be overcome. In this first section of this chapter, I will further develop solutions to create an expanded version of PF. These proposals primarily aim to set out the necessary conditions for making PF empirically testable. However, in my view, they will also contribute to PF's general strength as a theory of mind. The second section of this chapter builds on the additions to PF and describes the characteristics that a future neurotechnology should have to be suited for usage in an empirical test of PF like the Chip Test. These sections together formalize my contributions to the main aim of this thesis, which is to analyze the necessary conditions for empirically testing functionalism using an experimental philosophy of neurotechnology.

One aspect in which PF could be improved relates to how it defines functions. As mentioned in the previous chapter, PF is unclear about the level of detail at which functions should be defined. Because of this, a critic of the Chip Test of PF could explain away a partial replacement that distorts mental states: they can always argue that the relevant function has not been substituted in a way that is detailed enough. In this way, this version of functionalism is a hypothesis that cannot be falsified.

Consequently, to establish PF as a falsifiable and therefore testable theory, it has to include a condition that restricts the level of detail at which a function is described. In existing literature on PF, the only characteristic mentioned of a function is that it should be defined according to the best explanation of the brain's functions according to cognitive science. However, this still leaves open many possibilities for a nonbiological replacement of the function. For the sake of the argument, let us take short-term memory as a function that is part of the best scientific theory of cognition. Imagine furthermore that we know at the scale of neurons exactly how short-term memory functions in one or a limited number of brains. We could define that function based on a precise description of how relevant neurons respond to stimuli. It is highly unlikely however that short-term memory is realized in the exact same way in different people's brain at this micro-level. A definition of the function should for example leave room for people with a better or worse short-term memory. So, the definition of a function

in PF should at least be generalized in a way that accounts for the variance in how the function is executed across humans. According to Wheeler (2010, p. 6), the conditions for functional equivalence in PF should be loosened further to account for functional variation between animals as well. Multiple biological studies illustrate that certain animal species have developed capacities which can also be found in humans using significantly different underlying biological structures and mechanisms. Wheeler argues that this is further evidence that multiple realizability is not dependent on highly detailed definitions of functions.

Furthermore, requiring a definition of a function at a micro-level could make multiple realizability theoretically impossible. As mentioned by Schneider & Mandik (2018, p. 315), carbon-based systems have different chemical properties compared to systems like neurotechnologies that are built on silicon. It might be that a too fine-grained definition of a function directly leads to a physical impossibility of multiple realizability - at least with respect to silicon-based technology which is currently almost ubiquitous (Norton, 2021). If this were the case, the adjusted version of PF would in principle allow for multiple realizability, while at the same time always be able to dismiss empirical test results. Since I want to study in my thesis what is necessary for functionalism to be empirically testable, I propose that PF to this end should not define functions on a micro-scale.

Concludingly, I argue that PF, to allow for multiple realizability and for the testability of this feature, should define a function on a meso-level. On the one hand, this is to be distinguished from the micro-level at which a function is described in terms of activity of neurons or even smaller particles. On the other hand, in line with Dennett (2001), I hold that one should watch out for functional minimalism and think that very general macro-level reproduction of functions in nonbiological systems lead to similar mental states. I take the meso-level in-between to mean that a description of functions is at a level of detail that allows for physical realization with material that is not based on carbon. At the same time, a meso-level description contains more detail than common folk psychological descriptions of functions. I argue that mentioning that a function should be defined at a meso-level improves clarity of PF as a theory, even though ‘meso-level’, based on this description, remains a broad concept. I mentioned before that the Chip Test involves replacing a function at different levels of detail, so therefore I think it is sufficient to not provide a stricter definition of what I consider a meso-level description of a function. I also do not consider a strict definition to be possible.

A second important aspect in which PF falls short regarding empirical testability is its explicit mention of so-called ‘cyborgs’. They are ‘cybernetic organisms’ that are partly biological,

partly nonbiological; a "hybrid of machine and organism" (Haraway, 2006). In the previous chapter, I argued that PF does not entail a prediction of the effect of becoming a cyborg on mental states. This potentially allows proponents of PF to avoid criticism that might result from a method like the Chip Test in the future. PF is currently indifferent regarding empirical test results, giving proponents the ability to argue that they never held that cyborgs could have similar mental states to fully biological or fully nonbiological agents. Yet, this is a weak move in my view, because I contend that this is a natural consequence of the theory rejecting the importance of physical realization.

Thus, I argue that PF should be dedicated to the hypothesis that mental states should not significantly differ when one or more functions of a person are executed using nonbiological means. If mental states only depend on the execution of functions, a functionally equivalent cyborg should have similar mental states as well. In this way, PF better allows for empirical testability. The mental states of cyborgs have been considered in literature that combines extended mind theory with functionalism to create 'extended functionalism' (see for example Wheeler, 2010). Although literature on this theory does not explicitly mention empirical testability, it can still be a useful contribution to this chapter. This demands further explanation.

Extended mind theory, according to the often cited paper on the topic by Clark & Chalmers (1998), holds that bearers of mental content are not always fully determined by the biological body. It is one of the 4E approaches to cognition mentioned before[5]. The theory states that, under certain conditions, factors outside the biological body can partly constitute the mind. In their paper, Clark & Chalmers argue that this is not a science-fiction scenario that might happen in the future, but instead they write that only one condition needs to be fulfilled. They call this condition the Parity Principle and it holds:

> *"If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is […] part of the cognitive process." (Clark & Chalmers, 1998, p. 8)*

[5] Within 4E, the theory is referred to as extended cognition. I use 'extended mind theory' and 'extended cognition' interchangeably, as is common in literature (Drayson, 2010, p. 10). In their original paper on the extended mind, Clark & Chalmers (1998, p. 8) also mention that extended cognition entails an extended mind, as shown in the quote above. It goes beyond the scope of this thesis to address objections to this like the ones mentioned by Drayson (2010).

So, they argue that we should decide whether something should be considered as part of cognition without taking its physical location and composition into account.

This extended mind theory is considered to be dependent on a functionalist theory of mind by a number of scholars (e.g., Adams & Aizawa, 2009; Clark, 2008; Wheeler, 2010), even though this is not explicitly mentioned in the original paper. Just like functionalism, extended mind theory holds that mental states depend on the way something functions, not on its internal constitution (Levin, 2021). Extended mind theory also allows a mental state to be multiply realized. This shows in the Parity Principle, which only considers functional requirements for something to be part of the mind, without referencing physical realization (Wheeler, 2010, p. 4). Thus, according to extended functionalism, extended mind theory relies on the functionalist characteristics of multiple realizability and a definition of the mind that is based on functions.

I argue that integrating extended mind theory into functionalism is especially relevant and even necessary to allow PF to be empirically testable. Extended mind theory at its core entails that the human mind is not limited to the biological brain but can also include nonbiological elements when they meet the right functional requirements. This means that extended mind theory fits the gap I identified earlier by specifying that a mind can consist of different parts, some of which are biological, some of which are not. Extended functionalism, as opposed to a functionalist theory that does not endorse extended mind theory, thus explicitly entails the possibility for a cyborg to have mental states similar to humans.

I consider integrating extended mind theory with PF especially fruitful compared to other functionalist theories, since this significantly weakens a major objection to extended mind theory. To multiple scholars, extended mind theory as described in the original paper is too liberal in its conditions for considering something to be part of cognition and therefore too liberal in ascribing mental states (Rupert 2004; Adams & Aizawa, 2009; Aizawa, 2010). The paper that originally proposed extended mind theory (Clark & Chalmers, 1998) took a hypothetical person with Alzheimer's disease called 'Otto' and his notebook as a case study of the theory. Otto uses his notebook as an external memory. Clark & Chalmers argue that one should see Otto's notebook as part of his cognition, because it performs the same function as human memory, it is a constant in his life, the information from the notebook is automatically available, and Otto automatically endorses the information.

Scholars like Rupert (2004) consider this reasoning to be counterintuitive. He writes, as I do, that extended mind theory should be accompanied by a psychofunctionalist theory of mind. Based on this, Rupert (2004, p. 423) argues that "as cognitive science currently describes its explanatory kinds, they are not likely to have realizations with external components".

According to him, a notebook does not function according to a scientific theory of memory and should therefore not be considered part of cognition. However, extended functionalism, as I have proposed it, is in my view able to deal with these objections to extended mind theory by proposing more chauvinist conditions: extended functionalism is more restrictive in what artifacts are considered to be part of cognition. I argue that PF imposes less counterintuitive and controversial conditions for considering a (neuro)technology as a substitute for a cognitive function than the version of extended mind theory which Rupert is opposed to.

Concludingly, I propose to make PF empirically testable through two adjustments. I argue that the theory should define functions at a meso-level. In that way, the level of detail of a functional definition is high enough to not only allow for multiple realizability in theory, but also in actual (future) practice. At the same time, functional definitions are not defined so broadly that essential features of a function are overlooked. Besides, I recommend that PF should endorse extended mind theory so that it explicitly leads to the hypothesis that functionally equivalent cyborgs will have comparable mental states. The final section of this chapter discusses features required for a technology to be used in an empirical test of PF.

Conditions for a neurotechnology to be suitable for the Chip Test

I have just described the adjustments to PF I consider necessary to make the theory empirically testable. Now, I will shortly formalize the conditions I argue should be fulfilled before a neurotechnology can be used in the Chip Test.

One, the neurotechnology should fully substitute a cognitive function. A partial substitution or restoration of a function does not provide sufficient evidence to allow for interpretation of phenomenal reports. In the previous chapter, I have described how a modern technology like a cognitive hippocampal prosthesis is insufficient for empirical testing for that reason. Most existing neurotechnologies aim to restore cognitive functions (Müller & Rotter, 2017; Prochazka et al., 2001), while the Chip Test requires a full substitution of a function that was executed by a person without a medical condition.

Two, the technology needs to substitute a function that is general enough for people to report on their experience of it. I estimate that such a function will align closely with descriptions of functions in folk psychology. This can mean that the neurotechnology has to execute multiple lower-level functions as they are posited in cognitive science when a single lower-level function cannot be distinguished clearly in experience. PF requires this connection to cognitive science in definitions of functions.

Finally, the neurotechnology should substitute a function at the right level of detail. In general, I argue that the level of detail at which a function is recreated should be as high as possible for optimal value of test results. Again, it is hard if not impossible to define a threshold condition for the required level of detail. Still, those planning to execute the Chip Test should be wary of functional minimalism. A restriction on the maximum level of detail at which a technology replicates a biological function is not necessary. A higher degree of functional equivalence will always contribute to the strength of the test's conclusions. This is different from the proposed adjustments to PF, since not mentioning a limit there would harm empirical testability.

With these requirements for neurotechnology used in the Chip Test and the proposed adjustments to PF, I consider empirically testing functionalism using the Chip Test a real possibility that could lead to valuable insights. Again, neurotechnology will need to be developed further before it fulfills the proposed requirements. Also, as mentioned before, an empirical test of PF can provide the strongest implications when mental states remain similar after partial replacement. This would make a better justified case for strengthening PF than a distortion of mental states would make for weakening PF. This is mainly because altered mental states can often be ascribed to a lack of technological sophistication. Still, I am convinced that this thesis provides useful recommendations for connecting functionalism to concrete phenomena where it might otherwise be considered as a mere abstract theory of mind.

Conclusion

The aim of this thesis has been to investigate the necessary conditions for testing functionalism using an experimental philosophy of neurotechnology. Generally, philosophical theories of mind remain abstract, but I have argued that this field might be expanded to provide empirically testable hypotheses. I explained that it seemed reasonable to investigate the empirical testability of a functionalist theory of mind, since functionalism typically allows for multiple realizability. I selected a particular version of psychofunctionalism, PF, which I considered the strongest functionalist theory that also leaves room for multiple realizability. PF, contrary to machine state functionalism, allows an agent to be in multiple states at once. Moreover, PF's grounding in scientific theory enables more nuanced descriptions of qualitative experiences than those provided by analytic functionalism. Also, PF, unlike analytic functionalism, does not need to rely on folk psychological concepts, which I estimate to typically be more inaccurate than scientific terms. The features just mentioned pertain to psychofunctionalist theories in general, but I further outlined my specific version of PF. This theory should recognize the possibility of attributing mental states to systems which contain minds as a response to the China brain and Chinese room objections. Lastly, I argued that PF should be a type of role functionalism, whereby functions are equated to their role in a system, instead of their physical realizers. Again, this accommodates multiple realizability.

Then, I proposed Schneider's Chip Test as a method for testing functionalism. This test aims to study consciousness by partially substituting a brain function of someone with a chip and consequently relying on their phenomenal reports to assess whether their mental states related to this function are similar. This chip could be a neural prosthesis, and I mentioned a cognitive hippocampal prosthesis as an illustrative example of a concrete technology which is already being developed. Executing the Chip Test in this way does not self-evidently lead to clear conclusions, however. I have addressed a range of different concerns. It can for example be questioned whether one can trust phenomenal reports of 'cyborgs'. To such concerns, I have responded that an agent can reasonably be assumed to have similarly trustworthy reports to humans if they are constituted in a largely similar way. This argument is used to address the philosophical problem of other minds. Furthermore, I have argued that results of the Chip Test might be hard to interpret, because a brain function might not be replicated at the right level of detail. Even though the results do not entail self-evident conclusions for this reason, I maintained that repeated Chip Tests at a high level of detail could provide significant insights.

However, the Chip Test is more suitable for increasing confidence in PF when the test succeeds than for decreasing confidence in PF when the test fails. Besides, I explained that the current state of neurotechnology is insufficient for replacing a cognitive function at an adequate level.

The analysis of the Chip Test I just described, concerned the methodology required for an empirical test of functionalism. In the final chapter, I used this analysis to formalize the necessary conditions for empirically testing functionalism. These conditions included proposed adjustments to PF as well as requirements for the neurotechnology used in a Chip Test. Regarding the adjustments to PF, I argued that the theory should make explicit that functions are to be defined at a meso-level, even though this level cannot be strictly demarcated. Additionally, PF should include an endorsement of extended mind theory. In that way, PF recognizes that a nonbiological artifact can be constitutive of the human mind. As a consequence, PF becomes explicitly committed to the hypothesis that humans with a partial replacement will have similar mental states, if that replacement functions equivalently. Finally, a neurotechnology used in a Chip Test should fully replace a function that is general enough for someone to report on. At the same time, the neurotechnology should replicate the *mechanics* of that general function at a high level of detail.

I argue that I have provided a comprehensive investigation regarding the necessary conditions for an empirical test of functionalism. The conclusions of this thesis contribute to the philosophy of mind by establishing a connection between a generally abstract theory of mind and empirical phenomena. I hold that I have in this way proposed an improvement to functionalist theory and psychofunctionalism in particular.

The results of the empirical test proposed regarding PF could furthermore be a valuable addition to the ethical analysis of neurotechnologies. If the Chip Test were to lead to new insights regarding PF's hypothesis of multiple realizability, this can influence the assessment of the ethical desirability of neurotechnologies that replace cognitive functions. For example, if phenomenal reports after partial replacement indicate distorted mental states related to the function executed by neurotechnology, this is something to be taken into consideration in subsequent neurotechnology research.

Nonetheless, certain limitations of this thesis should also be recognized. I explained in chapter 4 that the proposed Chip Test does not provide strong, unambiguous proof for or against PF. For example, it can be hard to tell whether a cognitive function is substituted at the right level of detail by the neurotechnology. If one's mental states are distorted, but the neurotechnology only replicates a function in a general sense, it is likely that one cannot draw

any conclusions. Besides, it can be hard to assess whether a full cognitive function at the level of folk psychological terms is substituted, which I consider to be necessary for phenomenal reporting.

Regarding the technological side of the thesis, this project has not been focused on researching the current state of neurotechnology in great detail. It could be that recently technologies have already emerged that better fit the criteria I have proposed for an empirical test. However, this would presumably only result in increased urgency to empirically test functionalism in the near future and not directly lead to different findings in relation to my main research aim. It could be true that neurotechnologies will never be able to replicate a cognitive function in the right manner. Still, although we do not know when this will happen, I maintain that based on recent rapid developments in this research area, neurotechnologies will eventually be sophisticated enough to be used in an empirical test of functionalism.

In the upcoming years, I argue that further research is necessary regarding two aspects of this thesis in particular. Firstly, newly developed neurotechnologies should be investigated in the light of their suitability for an empirical test of functionalism. Especially in respect to the ethical implications of the functionalist hypothesis regarding partial replacement being false, it is crucial to perform an empirical test as early as possible. Secondly, this process can be accelerated by reviewing studies on the impact of existing neurotechnologies on one's mental states. These might not explicitly consider the role of philosophy of mind in this impact assessment, but nevertheless indicate what technologies and mental signs to look out for in an empirical test of functionalism.

References

Adams, F., & Aizawa, K. (2009). Why the Mind is Still in the Head. In M. Aydede & P. Robbins (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 78–95). Cambridge: Cambridge University Press.

Adams, F., & Aizawa, K. (2010). The value of cognitivism in thinking about extended cognition. *Phenomenology and the Cognitive Sciences*, *9*(4), 579–603. https://doi.org/10.1007/s11097-010-9184-9

Aizawa, K. (2010). The coupling-constitution fallacy revisited. *Cognitive Systems Research*, *11*(4), 332–342. https://doi.org/10.1016/j.cogsys.2010.07.001

Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/

Berger, T. W., Song, D., Chan, R. H. M., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A., & Granacki, J. J. (2012). A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *20*(2), 198–211. https://doi.org/10.1109/TNSRE.2012.2189133

Block, N. (2002). The Harder Problem of Consciousness. *The Journal of Philosophy*, *99*(8), 391. https://doi.org/10.2307/3655621

Block, N. (2007). *Consciousness, Function, and Representation: Collected Papers*. Bradford.

Block, N. (2013). Troubles with Functionalism. In *Readings in Philosophy of Psychology* (pp. 268–306). Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c31

Block, N., & Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *The Philosophical Review*, *108*(1), 1–46.

Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., & Guenther, F. H. (2010). Brain–computer interfaces for speech communication. *Speech Communication*, *52*(4), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

Carlson, N. R. (2013). *Physiology of behavior* (11th ed.). Pearson.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. OUP USA.

Chalmers, D. J. (2016). The Singularity: A Philosophical Analysis. In S. Schneider (Ed.),

Science Fiction and Philosophy (pp. 171–224). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118922590.ch16

Churchland, P. M. (1981). Eliminative Materialism and Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67–90. https://doi.org/10.5840/jphil198178268

Churchland, P. M. (2013). *Matter and Consciousness, third edition*. MIT Press.

Cinel, C., Valeriani, D., & Poli, R. (2019). Neurotechnologies for Human Cognitive Augmentation: Current State of the Art and Future Prospects. *Frontiers in Human Neuroscience*, *13*, 13. https://doi.org/10.3389/fnhum.2019.00013

Clark, A. (2008). Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind? *Philosophy and Phenomenological Research*, *76*(1), 37–59. https://doi.org/10.1111/j.1933-1592.2007.00114.x

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Costandi, M. (2016). *Neuroplasticity*. MIT Press.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1–2), 221–237. https://doi.org/10.1016/S0010-0277(00)00130-X

Dennett, D. (2018). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company.

Descartes, R. (2008). *Meditations on first philosophy: With selections from the objections and replies.* (M. Moriarty, Trans.). Oxford University Press. (Original work published 1641)

Drayson, Z. (2010). Extended cognition and the metaphysics of mind. *Cognitive Systems Research*, *11*(4), 367–377. https://doi.org/10.1016/j.cogsys.2010.05.002

Gennaro, R. J. (n.d.). Consciousness. *Internet Encyclopedia of Philosophy*. Retrieved May 25, 2022, from https://iep.utm.edu/consciousness/

Graham, G. (2019). Behaviorism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/behaviorism/

Hampson, R. E., Song, D., Robinson, B. S., Fetterhoff, D., Dakos, A. S., Roeder, B. M., She, X., Wicks, R. T., Witcher, M. R., Couture, D. E., Laxton, A. W., Munger-Clary, H., Popli, G., Sollman, M. J., Whitlow, C. T., Marmarelis, V. Z., Berger, T. W., & Deadwyler, S. A. (2018). Developing a hippocampal neural prosthetic to facilitate human memory encoding and recall. *Journal of Neural Engineering*, *15*(3), 036014. https://doi.org/10.1088/1741-2552/aaaed7

Haraway, D. (2006). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in

the Late 20th Century. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds.), *The International Handbook of Virtual Learning Environments* (pp. 117–158). Springer Netherlands. https://doi.org/10.1007/978-1-4020-3803-7_4

Haselager, P., Mecacci, G., & Wolkenstein, A. (2021). Can BCIs Enlighten the Concept of Agency? A Plea for an Experimental Philosophy of Neurotechnology. In O. Friedrich, A. Wolkenstein, C. Bublitz, R. J. Jox, & E. Racine (Eds.), *Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications* (pp. 55–68). Springer International Publishing. https://doi.org/10.1007/978-3-030-64590-8_5

Heil, J. (1998). *Philosophy of mind: A contemporary introduction*. Routledge.

Jacoby, H. (1989). Empirical Functionalism and Conceivability Arguments. *Philosophical Psychology*, *2*(3), 271–282. https://doi.org/10.1080/09515088908572979

Jaworski, W. (n.d.). Mind and Multiple Realizability. *Internet Encyclopedia of Philosophy*. Retrieved April 23, 2022, from https://iep.utm.edu/mult-rea/

Kamiński, J. (2017). Intermediate-Term Memory as a Bridge between Working and Long-Term Memory. *Journal of Neuroscience*, *37*(20), 5045–5047. https://doi.org/10.1523/JNEUROSCI.0604-17.2017

Knobe, J., & Nichols, S. (2017). Experimental Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/

Levin, J. (2021). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/functionalism/

Melnyk, A. (1994). Inference to the best explanation and other minds. *Australasian Journal of Philosophy*, *72*(4), 482–491. https://doi.org/10.1080/00048409412346281

Menary, R. (2010). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences*, *9*(4), 459–463. https://doi.org/10.1007/s11097-010-9187-6

Merriam-Webster. (n.d.). Consciousness. In *Merriam-Webster.com dictionary*. Retrieved May 25, 2022, from https://www.merriam-webster.com/dictionary/consciousness

Müller, O., & Rotter, S. (2017). Neurotechnology: Current Developments and Ethical Issues. *Frontiers in Systems Neuroscience*, *11*, 93. https://doi.org/10.3389/fnsys.2017.00093

Nagel, T. (2013). What Is It Like to Be a Bat? In *Readings in Philosophy of Psychology* (pp. 159–168). Harvard University Press.

https://doi.org/10.4159/harvard.9780674594623.c15

Newen, A., Bruin, L. D., & Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford University Press.

Nordmann, A., & Rip, A. (2009). Mind the gap revisited. *Nature Nanotechnology*, *4*(5), 273–274. https://doi.org/10.1038/nnano.2009.26

Norton, M. G. (2021). Silicon—The Material of Information. In M. G. Norton (Ed.), *Ten Materials That Shaped Our World* (pp. 177–195). Springer International Publishing. https://doi.org/10.1007/978-3-030-75213-2_11

Plato. (1997). Phaedo. In J. M. Cooper & D. S. Hutchinson (Eds.), *Complete works* (pp. 49–100). Hackett Pub. (Original work published ca. 390-385 B.C.E.)

Polger, T. W., & Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford University Press.

Popper, K. (2014). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.

Potter, S. M. (2010). Closing the loop between neurons and neurotechnology. *Frontiers in Neuroscience*, *4*. https://doi.org/10.3389/fnins.2010.00015

Prochazka, A., Mushahwar, V. K., & McCreery, D. B. (2001). Neural prostheses. *The Journal of Physiology*, *533*(1), 99–109. https://doi.org/10.1111/j.1469-7793.2001.0099b.x

Putnam, H. (1974). Meaning and Reference. *The Journal of Philosophy*, *70*(19), 699–711. https://doi.org/10.2307/2025079

Putnam, H. (1975a). Minds and Machines. In *Mind, Language and Reality: Philosophical Papers, Volume 2* (pp. 362–385). Cambridge University Press.

Putnam, H. (1975b). The Nature of Mental States. In *Mind, Language and Reality: Philosophical Papers, Volume 2* (pp. 429–440). Cambridge University Press.

Roelfsema, P. R., Denys, D., & Klink, P. C. (2018). Mind Reading and Writing: The Future of Neurotechnology. *Trends in Cognitive Sciences*, *22*(7), 598–610. https://doi.org/10.1016/j.tics.2018.04.001

Rowlands, M. J. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press.

Rupert, R. D. (2004). Challenges to the Hypothesis of Extended Cognition. *The Journal of Philosophy*, *101*(8), 389–428. https://doi.org/10.5840/jphil2004101826

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press. https://doi.org/10.1515/9780691197777

Schneider, S., & Mandik, P. (2018). How Philosophy of Mind Can Shape the Future. In A.

Kind (Ed.), *Philosophy of Mind in the Twentieth and Twenty-first Centuries* (pp. 303–319). https://philarchive.org/rec/SCHHPO-8

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

Shagrir, O. (2005). The Rise and Fall of Computational Functionalism. In Y. Ben-Menahem (Ed.), *Hilary Putnam* (pp. 220–250). Cambridge University Press.

Smart, J. (1959). Sensations and brain processes. *Philosophical Review*, *68*(April), 141–156.

Susan, S., & Mandik, P. (2018). How Philosophy of Mind Can Shape the Future. In A. Kind, *Philosophy of Mind in the 20th and 21th Century*. Routledge.

Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and Preattentive Processing. *The American Journal of Psychology*, *105*(2), 341. https://doi.org/10.2307/1423032

Trigg, J., & Kalish, M. (2011). Explaining How the Mind Works: On the Relation Between Cognitive Science and Philosophy. *Topics in Cognitive Science*, *3*(2), 399–424. https://doi.org/10.1111/j.1756-8765.2011.01142.x

Udell, D. B., & Schwitzgebel, E. (2021). Susan Schneider's Proposed Tests for AI Consciousness. *Journal of Consciousness Studies*, *28*(5–6), 121–144.

Van Gulick, R. (2021). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/consciousness/

Vázquez-Guardado, A., Yang, Y., Bandodkar, A. J., & Rogers, J. A. (2020). Recent advances in neurotechnologies with broad potential for neuroscience research. *Nature Neuroscience*, *23*(12), 1522–1536. https://doi.org/10.1038/s41593-020-00739-8

Wheeler, M. (2010). In Defense of Extend Functionalism. In R. Menary (Ed.), *The Extended Mind* (pp. 245–270). MIT Press.

Yablo, S. (2000). Textbook Kripkeanism and the Open Texture of Concepts. *Pacific Philosophical Quarterly*, *81*(1), 98–122. https://doi.org/10.1111/1468-0114.00097