Analytic Functionalism and the Ability Hypothesis versus Inversion Arguments

Author: Jasper van Boven

jsprvnbvn@gmail.com

Master Thesis Philosophy (60 EC)

Specialization: History and Philosophy of the Sciences

Supervisor: Dr. M.A. Lipman

Second Reader: Dr. M.S. van der Schaar

Leiden University

Table of contents

Introduction

If we were to ever create a perfect theory of mind, one of its central elements must be a convincing treatment of subjective experience. The concept of subjective experience is deeply problematic in itself: for many types of experiences, it seems impossible to convey all their qualitative properties to others. We cannot express the painful experience of breaking a bone in words, and no one becomes acquainted with the smell of vanilla through vivid descriptions. The trouble of communicating the qualities of experiences lies at the root of one of the main problems in the philosophy of mind: a complete description of the mind would surely contain a convincing account of subjective experiences, and our limited access to these experiences may seem to prevent such an account. There are various views on whether a satisfying solution to this problem is possible, but it is clear that any plausible theory of mind must address it in some way.

It is possible to roughly divide the viewpoints on the problem of subjective experience in two (very broad) groups. In one group, we have the philosophers who believe that we are able to give a plausible complete description of the subjective elements of experience. This perspective is often based on some version of physicalism or functionalism, and its supporters often attempt to characterize subjective experience in a way that allows for objective explanation based on psychological and neurological research.[1] Philosophers in the other group think that this approach to subjective experience is inadequate, and that there are at least *some* qualitative properties of experience that cannot be captured in objective terms.[2]

Philosophers like John Smart (1959), David Lewis (1966) and David Armstrong (1968) belong in the first group. They argue that since developments in natural sciences give us a plausible picture of so many phenomena, it is unreasonable to say that the mind cannot be properly captured in scientific terms. In an attempt to create an account of mental states that is in line with findings from the physical sciences, they started the development of a strand of theories of mind that is now known as *analytic functionalism*. Analytic functionalists attempt to create *topic-neutral analyses* of mental concepts:

[1] Examples of this view can be found in Smart (1959), Lewis (1966 & 1980), Armstrong (1968 & 1980), Churchland (1986 & 2002), Harman (1990), Tye (1994), Dennett (1991 & 2017).
[2] Examples of this view can be found in Broad (1925), Nagel (1974), Jackson (1982), Block (1980c, 1990 & 1996), Chalmers (1996 & 2010), Levine (2001).

analyses that are neutral about the physical or mental status of concepts.[3] This means that the descriptions of sensations do not contain terms that require us to accept or reject some sort of independent mental aspect of the world. By analyzing common-sense mental concepts (such as "feeling pain" or "being hungry") and determining their functional role, analytic functionalists attempt to find a place for folk-psychological concepts in a scientific framework.

There are philosophers who think that functionalist theories are too reductionist, and fail to account for the subjective aspects of experience in a proper way. Thomas Nagel (1974) argues that since the subjective properties of experience should be the central topic of scientific inquiry into the mental, it would make no sense to attempt a more objective description by taking the subjective elements out of experience. Frank Jackson's Knowledge Argument (1982) supports the view that a complete collection of all the physical knowledge would still leave some knowledge out, and this would mean that there must be some aspect of experience that analytic functionalism leaves out.

If we allow the notion that there is some type of irreducible qualitative aspect in subjective experience, we open the door for a group of arguments known as *inversion arguments*. Inversion arguments are arguments based on the intuition that even if two people are perfectly identical from a functional perspective, their subjective experiences could differ. This is usually illustrated with the concept of inverted color-experiences: it seems to be possible that when two people are presented with a blue object on the left and a yellow object on the right, person A's qualitative color-experience while looking at the left object is identical to person B's color-experience while looking at the right object (and vice versa). This difference would not appear in a functional description of these color-experiences, and this supposedly shows that there is some aspect of mental life that functionalist theories fail to properly capture.

While the specific potential failure to capture color-inversion in functional terms may be a weakness that the functionalist would be prepared to acknowledge, the possibility of this inversion would have deeper implications for functionalist theories. The failure to capture subjective differences between color-experiences would potentially also pose a problem for other types of experience. While a person's complete

[3] Block, 1980b: 179

functional description could suggest at state of perfect well-being, their actual experience may be one of deep pain. The fields of psychology and neuroscience would not be able to provide reliable knowledge about subjective experience: the characterization of experience used in those fields of research only captures the behavioral or functional aspects of the mind, and their claims only capture qualitative experience under the assumption that it is tied to those aspects.

It is clear that if we want a reliable and complete functionalist theory, it must be immune to inversion arguments. I think that inversion arguments result from a fundamentally flawed perspective on the qualitative properties of experience, and this paper is an attempt to argue for what I think is a more plausible perspective. I will defend the thesis that *analytic functionalism provides an appropriate framework for a plausible and complete analysis of mental states.* A defense of this thesis potentially requires a rejection of some deep intuitions regarding our subjective experiences, and it is therefore unlikely that many of its opponents will suddenly be swayed by this text. But despite the problems that affect any text concerning fundamental intuitions, I think that it is possible to make a relatively convincing case for the strength of analytic functionalism.

The main text of this paper has four parts. In the first part, I will provide a detailed account of analytic functionalism. This will start with a short reflection on the behaviorist roots of analytic functionalism, after which I will move on to a description of Smart's (1959) early functionalist theory and Lewis's (1966) stronger version. I will also consider some further developments that Lewis (1980) made, along with some of Armstrong's (1980) reflections on our conscious awareness of mental states.

The second part will start with a short description of the problems that Nagel (1974) and Jackson (1982) presented. While Nagel and Jackson only present somewhat general comments on the problems that a functionalist needs to answer for, I think that they provide clear examples of the underlying convictions that lead to opposition to functionalism. I will then present two versions of inversion arguments. The first version is a relatively simple one, which I will describe based on Shoemaker (1982). The second version is the more complicated argument from Inverted Earth, which was presented by Block (1990).

In the third part, I will propose a potential solution for the problems that inversion arguments pose. This solution will rely on accepting Laurence Nemirow (1990) and Lewis's (1999) so-called *ability hypothesis*, which roughly holds that our acquaintance with qualitative aspects of experiences lies in the ability to remember/imagine/recognize those qualitative aspects. I will first give a detailed explanation of the ability hypothesis, and then briefly show how this answers the problems that Nagel and Jackson presented. I will then argue that accepting the ability hypothesis gives us a characterization of qualitative experience that enables us to reject inversion arguments, and that these arguments can be shown to rely on implausible assumptions. Nemirow and Lewis argued that the ability hypothesis provided a satisfying answer to the issues that Nagel and Jackson saw. I intend to show that accepting the ability hypothesis also gives us a satisfying answer to the issues that the inversion arguments by Shoemaker and Block present.

In the fourth part, I will provide additional reasons for accepting the ability hypothesis. I will argue that acceptance of the ability hypothesis opens up the way for an account of qualitative experiences that is more plausible than alternatives. Dennett (1990) argues that traditional characterizations of qualitative experiences are often deeply problematic. I will show that the ability hypothesis underwrites a notion of qualitative experience that is less problematic, and not affected by Dennett's arguments against the traditional notion. I will then use texts by Shoemaker (1975), Quine (1969) and Dennett (2017) as a basis to reflect on the notion of similarity of qualitative experience, and then argue that there is a role for the ability hypothesis in a plausible account of qualitative experience.

The conclusion of the paper will be that a combination of analytic functionalism (AF) and the ability hypothesis (AH) gives us a plausible starting point for a theory of mind. I think that AF+AH is less problematic than alternative views, and that arguments against it often rely on intuitions that have far more troubling implications themselves. While I will not give a definitive proof for the success of AF+AH, I will show that combining AF and AH gives us a version of functionalism that is more plausible than many potential alternative accounts of the mental. This conclusion addresses fundamental issues in the philosophy of mind, and it may help towards strengthening the position of analytic functionalism.

Part I

An Overview of Analytic Functionalism

The behaviorist roots of functionalism

For a complete picture of analytic functionalism, it is useful to take a look at its behaviorist roots. The term ´behaviorism´ was introduced in 1913 by John B. Watson, and referred to the view that the field of psychology should be a "…purely objective experimental branch of natural science" that fully characterizes mental processes in terms of their behavioral expressions.[4] A main element in behaviorism is the view that psychological descriptions should not rely on references to 'inner' processes of the mind.[5] Behaviorism holds that the mind is not some mysterious internal substance that affects behavior: mental processes should be identified through their behavioral manifestations.[6] This view fits well with materialist accounts of the mind: outward behavior can be described in physico-chemical terms, meaning that successful identification of mental processes with outward behavior would then seem to allow for description of mental processes in physico-chemical terms.[7]

In *The Concept of Mind* (1949), Gilbert Ryle presents an account of the mind that fits well with behaviorism. Ryle attacks what he calls the 'official doctrine': the dualist view on the nature of minds that is very popular among both laymen and theorists (e.g. philosophers, psychologists, religious teachers).[8] He characterizes this view as follows:[9]

> -Every human being (perhaps with some exceptions, like very young children) has both a mind and a body.
>
> -Mind and body are usually harnessed together, but the mind could exist independent of the body.

[4] Watson, 1913: 248
[5] Armstrong, 1980: 193
[6] Ibid.
[7] Ibid.
[8] Ryle, 1949: 1
[9] Ibid.

-The body exists in space and is subject to mechanical laws. Its behavior is public, and can be inspected by observers.

-The mind does not exist in space, and is not bound by mechanical laws. Direct acquaintance with the processes of the mind is private: only the person whose mind it is can monitor at least some (and perhaps even all) mental states.

In Ryle's own account of the mental, reference to any sort of hidden mental substance is avoided. Mental processes are instead seen as an aspect of our behavior.[10] Anger does not cause aggressive behavior, it *is* the aggressive behavior.[11] Thoughts are not inner processes that could produce behavior, they are expressions of behavior.[12] Feelings do not mysteriously affect behavior: they are agitations that can be observed in our behavior.[13]

While mental processes can be identified based on observed behavior, a theory of mind should also be able to account for mental processes that do not directly affect our behavior. To account for these processes, Ryle uses the notion of *dispositions.* A disposition is the tendency of a thing to behave in a certain way under certain circumstances. Ryle uses glass as an example: to say that glass is brittle does not mean that it will ever break, it only means that it would easily break if a relatively small amount of force were ever applied to it.[14] The fact that the glass would break easily does not mean that glass is in the state of 'being brittle': it only means that it will change to a particular state (the state of 'broken') under certain circumstances.

If the notion of dispositions is applied to the mind, it is possible to create an account of mental states that do not directly produce behavior. A behaviorist could argue that although a person did not behave in a certain way, that person was disposed to behave in a certain way.[15] A person can be said to be angry without showing any signs of anger, if we know that there are specific events or circumstances that would suddenly result in them showing angry behavior.[16] The fact that the angry behavior was not

[10] Armstrong, 1980: 193
[11] Ibid.
[12] Ibid.
[13] Ryle, 1949: 91
[14] Ibid., p. 32
[15] Armstrong, 1980: 194
[16] Ibid.

actually produced does not lead the behaviorist to the conclusion that the person is not angry: the possibility of angry behavior is acknowledged, so there is still room for mental states that are not directly accompanied by actualized behavior.[17]

While the introduction of dispositions allows the behaviorist to acknowledge unobservable mental states, it does not open up the way for a complete behaviorist theory of mind. A science fiction example created by Hilary Putnam shows us that a behaviorist notion of mental states is simply too limited.[18] Putnam asks us to imagine a community where all the adults are able to successfully suppress all involuntary pain-behavior. They feel pain just like we do, but they have learned to hide the accompanying behavior for ideological reasons. They do not even speak about pain, as that would disgrace their family.[19]

If the community described in Putnam's example would indeed exist, then a behaviorist theory would fail to describe the mental aspects of its adult members. The behavior of the members would likely lead us to think that the members feel no pain. And even if a behaviorist would say that the members *could* still feel pain, their theory would not be able to determine whether they actually *do* feel pain or not. Since a complete theory of mind would have to account for all significant aspects of mental life, behaviorism's failure to determine the presence of fundamental mental processes (like pain) is a clear weakness.

While behaviorism can address some of the problems in the official doctrine, Putnam's scenario shows that a behaviorist analysis of the mental is too limited for a complete description of the mind. Even so, the promising parts of behaviorism play a role in the development of the type of functionalism that I will defend. Smart, Lewis, and Armstrong consider themselves descendants of behaviorists,[20] and I will now give an overview of the central points in their theories.

[17] Ibid.
[18] Putnam, 1980: 29
[19] Ibid, p. 30
[20] Block, 1980a: 175

Functionalism and the Identity Theory

When Smart (1959) laid the foundations for his version of functionalism, he did so for reasons of parsimony and simplicity.[21] Science suggests that organisms consist of nothing more that complex arrangements of physical constituents, and that future developments may allow us to describe behavior in fully mechanistic terms.[22] Non-physical sensations would be what Smart calls *nomological danglers:* things that fall outside the normal laws and explanations, and seem to occur according to some specific independent law.[23] It may seem possible that scientific developments will provide new laws to account for these nomological danglers, but this is very unlikely.[24] While the future will surely bring new discoveries of laws, these laws will likely describe very simple constituents. New laws that would describe processes consisting of billions of neurons would be unlike any other laws in science, and Smart is cautious of any argument that would rely on the existence of this complicated type of law.[25]

Smart argues that sensations are simply brain-processes. He defends the thesis that "…in so far as a sensation statement is a report of something, that something is in fact a brain process."[26] This does not mean that we can translate sensation statements to brain-process statements, or that they have the same logic as brain-process statements.[27] Smart says that sensations are brain-processes in the strict sense of identity: sensations are not only spatially or temporally continuous with brain-processes, they are the same four-dimensional object as the brain-process.[28] The hypothesis that experiences or sensations are the same as physical brain-processes is the basis for the so-called *identity theory*.

There are various objections to the idea that sensations statements refer to processes of the brain. The first issue that Smart addresses is that a person without any knowledge about brain-processes is still capable of reporting on their sensations.[29]

---

[21] Smart, 1959: 141, 155
[22] Ibid., p. 143
[23] Ibid.
[24] Ibid.
[25] Ibid.
[26] Ibid., p. 145
[27] Ibid.
[28] Ibid.
[29] Ibid., p. 146

Their sensation statements seemingly cannot refer to brain-processes, as the person would then refer to something that they know nothing about. Smart says that this is not actually a problem, because we can say something about A even if we do not know that "A is identical with B".[30] It can be said that a person reports on a brain-process if they say "I see a blue sky", even if the person has no idea that they are reporting on a brain-process.

Another issue is that sensations such as seeing color may seem to have qualities that are over and above the physical aspects of brain-processes.[31] Smart illustrates his solution to this problem by focusing on the concept of a secondary quality such as color. In Smart's view on color, a statement such as "This object is red" roughly means that a normal percipient would easily be able to pick the object out of a group of green things but not out of a group of red things.[32] This is usually possible thanks to the ability to perceive the color red, but a colorblind person would also be able to make this statement (for example, by relying on other people who tell them that the object is red).[33] This account of color suggests that color is an unimportant quality for physics, as the ability (or inability) of a complex neurophysiological mechanism to discriminate colors is unlikely to correspond to simple distinctions in nature.[34] This means that our color-sensations are not the result of exposure to some sort of intrinsic quality in nature.

Smart suggests that when a person says "I see a yellowish-orange after-image", this statement should be interpreted as something like: "*There is something going on which is like what is going on when I have my eyes open, am awake, and there is an orange illuminated in front of me, that is, when I really see an orange*." (Smart, 1959: 149). An important thing that Smart notes is that the sentence *"There is something going on which is like what is going on when…"* only contains topic-neutral words: there are no implications regarding the mental or physical status of the described experience.[35] This

[30] Ibid., p. 147
[31] Ibid., p. 148
[32] Ibid., p. 149
[33] Ibid.
[34] Ibid.
[35] Ibid.

partial sentence allows for descriptions that are neutral between dualism and materialism, as it only makes a claim about the similarity between two experiences.[36]

A stronger argument for functionalism

While Smart's claims support the widely held view that there is a strong relationship between sensations and physical processes, he has not found a definitive proof of this relation. He has only shown that it seems *highly likely* that sensations are physical processes, not that they must necessarily be so in the metaphysical sense. Saul Kripke says that in order for the identity theory to be proven, it must be shown that it is impossible to imagine that a specific sensation (or lack of sensation) does not correlate with a specific brain-process(or lack thereof).[37] Kripke compares statements regarding the identity of mind and body with statements like "Heat is the motion of molecules." *Heat* and *the motion of molecules* are necessarily identical. It is simply impossible to imagine that heat and the motion of molecules are different things: any attempt to imagine this would rely on incorrectly using one of the designators (or both) to designate something different than the thing that it in fact designates. Kripke illustrates this with the example of trying to imagine that the table in front of you might have been made of ice: this seems possible, but we are actually imagining a different table (made of ice) in the position of the actual table.[38]

If we want a definitive proof of the identity theory, it must be shown that the relation between mental events and physical events is the same as the relation between heat and the motion of molecules. David Lewis argues for the existence of this type of relation between mental events and physical events in 'An Argument for the Identity Theory' (1966). He sees the causal role of an experience as its definitive characteristic, and since these causal roles (according to him) necessarily belong to certain physical states, the physical states possess the definitive characteristics of the experience and are therefore identical with it.[39]

[36] Ibid.
[37] Kripke, 1980: 144
[38] Ibid.
[39] Lewis, 1966: 17

Lewis illustrates his argument by comparing experiences with cylindrical combination locks for bicycle chains. The definitive characteristic of their state of being unlocked is its causal state: setting the combination is the cause of the unlocking, and doing so has the effect of putting the lock in a state where it will open when gently pulled.[40] Knowing this allows us to determine whether the lock is in the state of locked or the state of unlocked. But the states of being locked or unlocked are also certain physical states: the alignment of slotted discs in the lock has the causal role of being unlocked, as the slotted discs need to be put in a specific configuration in order for the lock to be unlocked. Therefore the causal state of being unlocked is identical with the physical state of being unlocked.[41] Lewis thinks that the same reasoning can also be applied to experiences.

There is a difference between an experience and the attribute of having that experience. The state itself *is* the state that occupies a certain definitive causal role, while the attribute of having that experience is the attribute of being in that state.[42] This distinction allows Lewis to refute the following objection:[43]

> -Attributes are only identical if they are predicated by synonymous expressions.
>
> -Experience-ascriptions are never synonymous with neural-state-descriptions.
>
> -Therefore, experience-ascriptions and neural-state-descriptions are never identical.

The distinction between the experience itself and the attribute of having that experience shows that this objection fails. The objection only shows that there is a difference between ascribing an experience to someone and describing their neural state, which Lewis agrees with.[44] It does not show that experiences and neural states are different, so we can still hold on to this view.[45]

---

[40] Ibid.
[41] Ibid., p. 18
[42] Ibid., p. 19
[43] Ibid.
[44] Ibid.
[45] Ibid.

Lewis says that the definitive causal role of any experience is "…expressible by a finite set of conditions that specify its typical causes and its typical effects under various circumstances."[46] While Lewis's description of experiences does not fit in behaviorism (as thorough behaviorist analyses deny this characterization of mental states, including experiences), the principle that experiences are defined by their causal roles is based on the behaviorist view that there is a necessary connection between an experience and its occasions and manifestations.[47]

A big difference between behaviorism and Lewis's theory is that Lewis's holds that experiences are something real: in his account, experiences are the "effects of their occasions and the causes of their manifestations".[48] A second difference between Lewis's theory and behaviorism is that Lewis claims that an experience can be caused by other experiences, and can cause other experiences itself.[49] This aspect of his theory is important, because it is required if we want to say that experiences are accessible through introspection: if we say that experience y is the experience of reflecting on experience x, then experience x must be such that it somehow is a causal factor for experience y. If there is nothing in experience x that allows the potential causing of other experiences, we would not be able to have the experience of reflecting on experience x.

A third difference is that in Lewis's theory, it is not the case that all causes and effects of all occurrences of the experiences are relevant in defining the experience.[50] It is acceptable to only use the causes and effects that typically determine the experience in order to define the experience, and exceptional cases where typical causes or effects are absent (or present while they typically are not) are not problematic.[51] Since behaviorism does not acknowledge the existence of an experience apart from its occasions and manifestations, a behaviorist definition cannot focus on typical causes or effects and instead has to be very complex (as there is no "typical occurrence" of an experience to fall back on).[52]

---

[46] Ibid.
[47] Ibid., p. 21
[48] Ibid.
[49] Ibid.
[50] Ibid., p. 22
[51] Ibid.
[52] Ibid.

Lewis thinks that while nonphysical phenomena may coexist with physical phenomena (ranging from perfectly correlating with them to being unrelated to them), they cannot have any causal power towards physical phenomena.[53] Since the behavioral manifestations of experiences always involve physical phenomena, they cannot be the result of something that is nonphysical.[54] Because these behavioral manifestations are among the typical effects that definitively belong to an experience, the experience cannot be the result of something that is not physical.[55] And even though it is not positively established that neural states are the definitive causal roles of experiences, there does not seem to be any other physical phenomena that could plausibly fit the role.[56]

How to account for unusual mental states

Even if we would accept the plausibility of Lewis's theory, there may be some instances of mental states that are more problematic. Lewis (1980) presents two problematic types of pain that may occur. The first is what he calls *mad pain*: a feeling of pain that is just like our usual notion of pain, but connected to unusual causes and unusual effects.[57] The second type is what he calls *Martian pain*: pain with causes and effects that are similar to those in the case of the average human being, while the physical instantiation has significant differences.

The two unusual cases must have a place in a good theory of mind. An identity theory that focuses purely on physical configuration can provide a proper description of mad pain: we could say that the madman is in pain because his brain is in a state of feeling pain, even though that brain state was reached by unusual causes and causes unusual effects.[58] But while such a theory could describe the pain of the madman, it would fail to capture the pain of the Martian: the Martian's feeling of pain would not be recognized by a theory that identifies pain with the usual human brain-configuration during pain.

[53] Ibid., p. 24
[54] Ibid.
[55] Ibid.
[56] Ibid.
[57] Lewis, 1980: 216
[58] Ibid., p. 217

A functionalist or behaviorist theory of mind could describe the pain of the Martian.[59] Since his pain has similar causes and similar effects to ours, the functionalist or behaviorist could say that all the relevant factors point towards the conclusion that the Martian does indeed have pain. But this theory would not be able to capture the pain of the madman: his feeling of pain is not related to any usual causes, and does not have the usual effects.[60]

Lewis thinks that the theories that he and Armstrong (1968) have proposed are able to account for both mad pain and Martian pain.[61] This is because he and Armstrong understand the concept of pain to be a nonrigid concept.[62] What this means is that while the concept of pain usually applies to a specific type of neural state, it could also have been applied to some other physical state if that state had a similar causal role.[63] The fact that we are in pain during certain neural states is contingent: the concept of pain could be applicable in many other possible cases, it just happens to apply to normal humans during a certain type of neural state.[64]

By saying that pain is some state that occupies a causal role for a population, we can account for both normal humans and Martians by saying that pain has the same causal role in both populations.[65] Pain is [a certain pattern of neurons firing] for the population of humans, and [the inflation of certain cavities] for the population of Martians.[66] But this physical difference is not problematic, as we can still say that the two different physical states embody the same causal state.

Lewis addresses the madman by saying that his theory allows for exceptional cases.[67] We have seen earlier that his theory holds for cases with *typical* causes and effects, and he repeats that statement.[68] The madman and his fellow madmen are the

---

[59] Ibid.
[60] Ibid.
[61] Ibid., p. 218
[62] Ibid.
[63] Ibid.
[64] Lewis notes that this does *not* mean that the concept of pain and our neural state of pain are contingently identical (1980: 218). It is contingently true that pain is a certain neural state, but it is necessary that pain is identical with itself and therefore also identical with the certain neural state.
[65] Lewis, 1966: 219.
[66] Ibid.
[67] Ibid.
[68] Ibid.

exceptional cases where the typical causes and effects do not apply. But they are still part of the human population.[69] Lewis say that we can still say that the madman is in pain, because he is in the same state as the others in his population when they are in pain.[70]

Awareness of mental states

A potential weakness for functionalist theories may be that they do not provide a clear role for our first-person awareness of our mental states. While functionalism may give us a satisfactory explanation of the behavior of others, it may not seem to account for everything that seems to be going on in our mind. In 'The Nature of Mind' (1980), Armstrong illustrates this with the example of automatism that may occur when people drive a car for a long time without a break: it is possible to drive a car for a long distance, and then suddenly realize that you have driven a long period without being aware of what you were doing. Clearly there must have been some unconscious mental processes going on; the car would have crashed otherwise.[71] But the mental activity during the ride was different than the mental activity that we have while consciously driving a car, as we are aware of our actions in the latter case.[72] The difference between conscious awareness and unconscious awareness seems significant in a description of the mind, so we must pay some attention to it.

Armstrong thinks that the problem of describing consciousness can be solved by defining consciousness (in the sense used in the driving-example) as "…perception or awareness of the state of our own mind."[73] We could then say that the driver in the automatic state is aware of the road, but not aware of his awareness of the road.[74] If this account is right, inner observation can be seen as similar to perception. Consciousness can be seen as awareness of inner states, similar to how perception is awareness of outer states.[75] Awareness of thoughts or emotions comes with the capacity to discriminate between different mental states.[76] This capacity can be expressed with

[69] Ibid.
[70] Ibid.
[71] Armstrong, 1980: 198
[72] Ibid.
[73] Ibid.
[74] Ibid., p. 199
[75] Ibid.
[76] Ibid.

words, but expressions are only results of the inner state and not the inner state itself.[77] If we regard consciousness to be awareness of inner states, we can capture it in functional terms: a state of consciousness can be seen as a causal state, and therefore fits into Lewis and Armstrong's functionalist accounts of the mental.

Part II

Problems for Analytic Functionalism

Potential issues for description of subjective experience

A main reason for the attractiveness of functionalism is that it enables us to make many objectives statements about mental states, regarding both our own mental states and those of others. But the broad scope of functional statements may lead to doubts about whether these statements can provide an exhaustive account of all the relevant features of the mind. In 'What is it Like to be a Bat?' (1974), Nagel argues that reductionist theories like those of Armstrong (1968) and Lewis (1966) completely fail to adequately address the relation between the body and consciousness. Nagel says that if a creature has consciousness, then there must be something that it *is like to be* that creature.[78] This 'what it is like'-aspect of consciousness is what he calls the subjective character of experience.[79] He thinks that this aspect is not analyzable in functional or causal terms, because he thinks that an analysis of this type would be logically compatible with the absence of subjective experience.[80]

Nagel gives a description of how objective descriptions are achieved in science. Many phenomena can be described from many points of view and through different perceptual systems.[81] A reductive explanation of experience is problematic, because our

[77] Ibid.
[78] Nagel, 1974: 166
[79] Ibid.
[80] Ibid., p. 167
[81] Ibid.

subjective viewpoint is precisely what we wish to describe.[82] Nagel says that our subjective viewpoint of what it is like to experience a thing is the most comprehensible, and a more objective description would be from a viewpoint that is actually farther away from the real nature of the phenomena of experience.[83]

A more focused version of Nagel's problem can be found in Jackson's Knowledge Argument. We have to imagine a brilliant scientist named Mary, who has always been locked in a black and white room and has to investigate the world from a black and white television.[84] This television provides her with every possible piece of physical information about the neurophysiology of vision: how the wave-length of light affects the retina, what goes on in the brain when we see certain colors, when and how our central nervous system produces sentences such as 'The sky is blue', etcetera.[85] It is plausible that all physical information could be received in this way. But if Mary ever leaves the room, it seems obvious that she will learn something new about visual experience when she is confronted with actual colors.[86] While she supposedly had all the physical information, she still obtained new information, implying that physical information is not all the information that there is.[87] We can see that Jackson holds the view that knowing what some experience is like is propositional knowledge, and that the only way to obtain this propositional knowledge is by having a certain experience.

The knowledge argument is a potential problem for analytic functionalism. It shows that our ability to talk about the mental states of others may not be that strong at all: we can regard color-experience as the result of some observation and as a causal element for various behaviors, but we may not be able to produce an exhaustive account of color-experience. Intuitively, there seems to be a difference between people who have had an experience and people who are only familiar with that experience through knowledge of its causes and effects.

[82] Ibid., p. 175
[83] Ibid.
[84] Jackson, 1982: 130
[85] Ibid.
[86] Ibid.
[87] Ibid.

Nagel and Jackson show some problems that functionalists need to address, but there may be an even more fundamental problem for functionalism. Nagel says that there is a sense in which we can say objective things about the experiences of other people, as long as our own point of view is similar enough.[88] Jackson seems to have a similar view: the Knowledge Argument is intended to show that Mary lacks certain knowledge of color-experience, but the circumstances under which she would gain this knowledge are clear. Their arguments do not account for the fact that knowledge of subjective experiences may be far more limited than they assume, and that our experiences may be far less similar (or completely different) to those of others (or our own at some other time) than we assume. This implication of the divide between subjective and objective knowledge can be illustrated with a look at the implications of inversion arguments.

Inversion Arguments

Before I turn to Shoemaker (1982) and Block's (1990) treatments of inversion arguments, I will first give a more general account of inversion arguments and the problems that they potentially pose. In a certain sense, it seems likely that the way that colors look is more or less the same between different people. Imagine looking at a blue sky, together with another person. From the way we usually discuss colors, it may seem obvious that you both see the same color (assuming that you both have relatively normal functioning optical systems). You assume that the other person's qualitative blue-experience is similar to yours, and we would be confused if they would suddenly genuinely say something like "the sky is bright yellow today".

While color-descriptions usually do not pose practical issues, it seems possible to intuitively imagine a sense in which the other person could see a completely different color. The fact that both you and the other person refer to the sky as 'blue' is only because you have both learned to refer to the color of the sky (and all things with a color similar to it) with the word 'blue'. The idea that the other person has the same color experiences as us may seem obvious because they use the same words to refer to these experiences as we do, but this actually only shows a non-qualitative similarity between you and the other person. When you use the words 'blue' and 'yellow', you associate

[88] Nagel, 1974: 172

these with your color-experiences of seeing blue and yellow. But what if these experiences are swapped for the other person? The color-experience that they have when looking at a blue sky could be the same color-experience that you have when you look at a lemon, and vice versa. The fact that we use the same language cannot solve the issue: the words that we use do not refer to our personal color-perception, but only to the sense in which color-perceptions are similar to all of us with average color-perception.

The thought that color-experiences can be inverted is often presented in the form of spectrum-inversion: the idea that there is a color spectrum that can be inversed (or perhaps shifted in some other way) between different people. If spectrum inversion is conceivable, it opens up the way for inversion arguments against functionalism.[89] A simple standard inversion argument looks something like this:

> -If spectrum inversion is conceivable, even the most detailed functional description of the mind cannot reliably tell us what kind of qualitative color-experience the subject has.
>
> -Qualitative color-experiences are a significant aspect of our mind.
>
> -Conclusion: If spectrum inversion is conceivable, functionalist theories are unable to provide a complete description of the mind.

While I think that analytic functionalism combined with the ability hypothesis will give us a plausible theory of mind, the potential conceivability of spectrum inversion would indeed be a significant obstacle. This means that in order to defend analytic functionalism, it must be shown that a plausible functionalist theory can treat the concepts involved in inversion arguments in such a way that spectrum inversion becomes inconceivable or at least highly implausible. I will now give a more detailed account of two inversion arguments and the concepts that are involved in them, based on texts by Shoemaker and Block.

[89] Shoemaker, 1982: 367

Shoemaker's treatment of spectrum inversion

One of the distinctions that Shoemaker illustrates in 'The Inverted Spectrum' (1982) is the distinction between *intra*subjective and *inter*subjective spectrum inversion. *Intra*subjective inversion refers to a systematic difference between a person's color-experience at a certain time and that same person's color-experience at a different time.[90] *Inter*subjective inversion refers to a systematic difference between the color-experiences of two different people when presented with the same external color-information.[91] The potential conceivability of intrasubjective inversion does not imply the conceivability of intersubjective inversion, but Shoemaker thinks that the conceivability of the former (if it is indeed conceivable) can possibly be used to argue for the plausibility of conceivability of the latter.[92]

For his argument, Shoemaker first argues that it seems plausible that intrasubjective inversion is possible.[93] Based on a passage from Wittgenstein's lecture notes (1968), Shoemaker argues that we can easily imagine how we could verify a change in a person's color-experience.[94] Wittgenstein gives the example of a person who says that their color-experience of red and blue have somehow switched.[95] The person says that glowing coals look cold, while the clear sky looks warm. Wittgenstein thinks that we would be inclined to say that the color-experiences were indeed switched.[96] Shoemaker agrees, and thinks that we could empirically determine whether the person has actually undergone this inversion.[97]

In Wittgenstein's example, we can determine that the person's qualitative experience is different because it can be empirically tested: we can confront the person with various colored objects and ask whether their color-experience is different than it was before. But if the other person has had different color-experiences as us from birth (or somehow forgot about his earlier color-experiences), we would not be able to tell that something was different between ourselves and the other person. If intrasubjective

[90] Shoemaker, 1982: 358
[91] Ibid., p. 357
[92] Ibid., p. 358
[93] Ibid., p. 359
[94] Ibid.
[95] Wittgenstein, 1968: 284
[96] Ibid.
[97] Shoemaker, 1982: 359

spectrum inversion is conceivable, it could lead to great skepticism about other minds: if the behavior of others does not tell us anything about their qualitative experiences, we may never know they are actually experiencing.[98] But Shoemaker thinks that the problem can be solved. To understand why, it is useful to first look at his distinction between *intentional* and *qualitative* content.

Shoemaker distinguishes between an object's *intentional* content and its *qualitative* content. He make this distinction in the context of a person who has undergone intrasubjective spectrum inversion, but has eventually accommodated to the change in his experience.[99] This person sees yellow in instances where he would previously see blue, but has made an effort to change his vocabulary: he now uses the term 'blue' to refer to the color of the sky, even though the color that he sees is the one that he would have referred to as 'yellow' before his spectrum inversion.[100]

The spectrum inversion has not changed the intentional content of the person's experience: the external color-information that he is confronted with remains the same, and he has learned to use the same words to refer to it as he did before.[101] But there has occurred a change in the qualitative aspect of his color-experiences: the color-experience that he has when looking at 'the blue sky' has a different qualitative aspect than before his inversion.[102] In that sense, his color-experience when presented with blue is the one that he previously would have had when presented with yellow. But while we can say that our descriptions of experiences refer to the intentional content of the experience, this still leaves problems for the qualitative aspect of experience. Even if your intentional content and behavior regarding some phenomenon is similar to other people, there may still be a radical difference between your qualitative experience and that of others.[103]

Shoemaker reflects on what he calls the *Frege-Schlick view* on qualia. Frege's (1956) view is that references to one's qualitative experience are only meaningful to the experiencer. Schlick's view is that we can only make meaningless statements about the

[98] Ibid., p. 364
[99] Ibid., p. 365
[100] Ibid.
[101] Ibid.
[102] Ibid.
[103] Ibid., p. 368

qualitative similarity of the experience of two subjects, as we cannot properly refer to these experiences.[104] Shoemaker says that if the Frege-Schlick view holds, it is unproblematic to talk about the similarity of our experiences and those of others.[105] This is because the F-S view holds that our descriptions only refer to the intentional content of experiences, and that comparisons between our experience and that of others therefore only involve the intentional content. Qualitative similarities affect one's behavior, and the beliefs of the person towards experiences.[106] But since the relations between a qualitative state and behavior do not hold intersubjectively, intersubjective inversion is not a relevant problem: it simply makes no sense to talk about the qualitative similarity between your qualitative experience and that of others.[107]

Shoemaker thinks that qualia (which he defines as "…features of sensory states in virtue of which they stand to one another in relationships of qualitative similarity and difference") can be defined in functional terms, by giving a functionalist account of what it means for qualia to be qualitatively similar.[108] A functionalist description would have to be that if qualia Q1 and Q2 are similar (to a certain degree), then the experience of Q1 and Q2 in a person will both have similar effects on the person's beliefs and behavior.[109] Two people can have similar qualia, though these may not cause a similar qualitative experience for these people: it only means that both qualia would cause a similar qualitative experience if they were to be instantiated in one of the people.[110] But while the qualitative experience that the similar qualia cause may be different for two people, the physical/functional instantiation of the qualia will be similar for both people.[111]

I think that there is a problem in Shoemaker's treatment of qualia and spectrum inversion. It is unclear to me how Shoemaker's characterization of qualia is helpful in the discussion about spectrum inversion. The problem that qualia originally presented for functionalist descriptions of mental states was that the same functionalist description of a mental state could potentially apply to multiple qualitatively different states. These states would appear identical from a purely functional description, and the fact that they

[104] Schlick, 1959: 93
[105] Shoemaker, 1982: 371
[106] Ibid., p. 370
[107] Ibid., p. 372
[108] Ibid.
[109] Ibid.
[110] Ibid.
[111] Ibid.

contained different qualia (or no qualia at all) was a significant factor that the functional description missed. If we focus on the supposed physical/functional instantiation of qualia, we are missing the problem that the usual conception of qualia presents. Shoemaker presents a conception of qualia that may be captured in functional terms, but inversion arguments seem to involve a conception of qualia that evades Shoemaker's treatment. We can say that qualia Q1 and Q2 are similar if they have similar effects on a subject's beliefs and behavior, but inversion arguments hold that two different qualia could have the same relations to a subject's beliefs and behavior. If we want to attack inversion arguments, we need to show why this is implausible.

Block's Inverted Earth

Block has proposed a unique version of spectrum inversion. He wants us to imagine an alternative version of Earth: Inverted Earth. Inverted Earth is exactly the same as our Earth, except for two things: everything on Inverted Earth has the complementary color of its color on Earth, and the hue-related vocabulary of Inverted Earth's citizens is inverted.[112] The grass on Inverted Earth is red, yet its residents would truthfully answer "green" when asked about the color of grass. We cannot say that one of the two Earths has the 'correct' terms for colors, in the same way that we cannot say that the English term "blue" is more or less correct than the French term "bleu". Inverted Earth residents are not incorrect when they say that "grass is green", as their use of the word "green" refers to the color that we refer to with the word "red". The difference between the usual inverted spectrum hypothesis and the Inverted Earth-scenario lies in the fact that the former is an attempt to show the conceivability of qualitative differences in cases with identical intentional content and functional descriptions. The Inverted Earth example will supposedly show the possibility of identical qualitative experiences in the case of different intentional content and functional descriptions.[113]

Block lets us imagine that we are knocked out by a mad scientist, who then puts color inversion lenses in our eyes and transports us to Inverted Earth.[114] When you wake up, there is no reason to suppose that you are on Inverted Earth at all. Your situation on Inverted Earth seems exactly the same as on Earth, because the lenses

[112] Block, 1990: 62
[113] Ibid., p. 62
[114] Ibid., p. 63

prevent you from seeing that the world around you is a different color than the world you know. This means that the qualitative content of your experience is identical to the one you would have had on Earth.[115]

According to Block, your intentional content on Inverted Earth is more problematic. Block thinks that your beliefs about intentional contents remain the same upon arriving at Inverted Earth.[116] Since all your color concepts are based on experiences on Earth, your claim that "grass is green" would be wrong on Inverted Earth: your conception of green is based on the color of grass on normal Earth, and grass is not that color on Inverted Earth. Your statement that "The sky is as blue as ever" would be incorrect, as the sky that you see on Inverted Earth is yellow and not blue. Your statement cannot refer to the color of the sky, as your lenses prevent you from seeing the actual color of the sky. This means that the intentional content of your experience does not correspond with the color of the sky of Inverted Earth. After a long time, your color-terms may come to refer to the colors on Inverted Earth. Once this has happened, your functional structure has changed: the effects that blue and yellow things have on your functional state have swapped.[117]

[115] Ibid.
[116] Ibid., p. 64
[117] Ibid.

Part III

A Potential Solution for the Problem of Spectrum Inversion

The ability hypothesis

Before I turn to proposing a solution for inversion problems with help of the ability hypothesis, I will first present the ability hypothesis in the context of the arguments that Nagel and Jackson presented. In 'Physicalism and the Cognitive Role of Acquaintance' (1990), Nemirow addresses a difficulty for physicalist theories of mind. Some philosophers (e.g. Feigl 1967, Nagel 1974) argue that physical information is objective, while sensory information is subjective. Feigl argues that sensory information is not knowledge. He says that a blind person does not lack knowledge about supposed qualities of sight, but only the knowledge that can be obtained through sight.[118] Feigl thinks that the knowledge that we obtain by sight does not necessarily have to be obtained through sight, and that it can also be obtained through other forms of observing. Nemirow thinks that Feigl underestimates the problem, because a blind person could not know what it *is like* to see.[119] Knowledge of experience can only be obtained through having the experience, and words such as "knowing", "discovering", and "remembering" are usually also used for the type of knowledge that we gain through experience.[120]

While Nemirow agrees with Nagel and Jackson's opposition towards Feigl, he thinks that their own accounts of the cognitive role of acquaintance are flawed.[121] He says that the knowledge argument only works if we assume that knowledge of what something is like must be knowledge of the way things are.[122] This assumption would allow us to say that since knowing what something is like is not based on physical theorizing, information about what something is like is not part of physical science.[123]

[118] Feigl, 1967: 68
[119] Nemirow, 1990: 491
[120] Ibid.
[121] Ibid., p. 492
[122] Ibid.
[123] Ibid.

But the assumption is problematic, because our vocabulary regarding knowledge also applies to abilities.[124]

There seems to be a correlation between knowing what something is like and knowing how to imagine it, which also fits well with Nagel's claims.[125] We cannot say that we can imagine what something is like, and at the same time not know what it would be like to experience the thing.[126] This strong correlation leads Nemirow to suggest what he calls the *ability equation*: knowing what it is like to have a certain experience is identical to knowing how to successfully imagine having that experience.[127]

The ability equation can be used to answer many problems. It would refute the knowledge argument, as it would show the falsity of the proposition that knowing what something is like is propositional knowledge.[128] It would also explain why it is appropriate to use vocabulary regarding knowledge (words such as "knowing" and discovering) when we talk about what an experience is like.[129] A third solution that it provides is the fact that we would not have to attribute subjectivity to certain experiences: failure to know what something is like is simply the lack of ability to imagine it, and there is no need to attribute some intrinsic subjectivity to those who do have the ability.[130] A fourth problem where it would help is the inexpressibility of knowing what something is like: instead of saying that experiences have inexpressible qualities (as opponents of physicalism would do), it allows us to say that knowing what something is like is only linguistically inexpressible because it is an ability.[131]

Even though replacing the concept of inexpressible qualities with linguistically inexpressible knowledge helps physicalism, physicalists still have to show why the 'what it is like'-knowledge is inexpressible.[132] Nemirow answers this using the example of color. The ability to imagine a color can only be communicated to someone who has the

[124] Ibid.
[125] Ibid.
[126] Ibid., p. 493
[127] Ibid.
[128] Ibid.
[129] Ibid.
[130] Ibid.
[131] Ibid., p. 494
[132] Ibid.

ability to perform an action by which they can visualize the color.[133] Usually only the following three actions are suitable: directly visualizing the color, remembering the visual experience of the color, or visualizing or remembering similar colors and interpolating.[134] When someone is able to perform one of these actions, we can instruct them to perform the other actions.[135] When someone is not able to perform any of these actions, we simply cannot tell them what seeing the intended color is like.[136]

An obstacle for the ability equation is the intuition that imagination grants direct access to universals.[137] What this intuition means is that our ability to visualize a certain color (such as red) may seem to be the ability to access some sort of 'raw redness that is an of intrinsic part of the world. Nemirow says that this is actually not the case: the ability of visualizing red is the ability to represent the particular perceptions of the color red.[138] Successful visualization allows us to compare colors and make statements about other colors. For example: when you witness a color x, you can imagine a darker shade y of that color and say "the color that I am witnessing now has a lighter shade than y".[139]

The imagining of a color has the same functional role as actually seeing the color, allowing us to make propositional statements about our imagined color in the same way that we do when actually presented with the color.[140] The same can be applied to the notion of pain: Nemirow gives the example of avoiding the dentist because we imagine that the pain that a dentist will inflict outweighs the benefits of the visit.[141] Our imagined pain functionally represents the expected actual pain in the future, and allows us to reason as if we were having the imagined experience.[142] Because we are able to use our imagination so effectively in our reasoning about actual experiences, it may seem that the imagined concepts (like color and pain) are actually universals with subjective

[133] Ibid., p. 493. "Communicating the ability to imagine a color" does not mean transmitting that ability through language, it refers to the ability to successfully tell others that you have the ability to imagine the intended color.
[134] Ibid., 494
[135] Ibid.
[136] Ibid.
[137] Ibid. p. 495
[138] Ibid.
[139] Ibid.
[140] Ibid., p. 496
[141] Ibid.
[142] Ibid.

qualities.[143] But this intuition is incorrect, as our successful imagining does not allow us to infer that we have direct access to the essential qualities of the experiences.[144]

Lewis (1999) further defends the ability equation, or as he presents it, the ability hypothesis. The ability hypothesis is very similar to Nemirow's ability equation, and Lewis endorses his view.[145] His own formulation is as follows: "The Ability Hypothesis says that knowing what an experience is like just *is* the possession of the abilities to remember, imagine, and recognize."[146] He uses the term hypothesis because it is presented as a more plausible alternative to a different hypothesis: the Hypothesis of Phenomenal Information. The Hypothesis of Phenomenal Information holds that there is such a thing as phenomenal information: nonphysical and irreducible information involved in an experience, that enables us to know what it is like to have that experience.[147] Acceptance of the ability hypothesis would imply rejection of the Hypothesis of Phenomenal Information: if we see the qualitative aspect of an experience as the ability to remember/imagine/recognize that experience, the supposedly irreducible information is simply information that can only be accessed by those who have had the experience. It is simply information that is accessed in a specific way, but the information itself is not of a completely different type.

How the ability hypothesis could solve the spectrum inversion problem

Nemirow and Lewis argue that the ability hypothesis can be used to refute the knowledge argument. I think that if we were to accept the ability hypothesis, it would also be possible to refute inversion arguments and thereby strengthen the position of analytic functionalism. I will reflect on the described types of inversion arguments, and show how they can be accounted for if we keep the ability hypothesis in mind. The first type is the standard type, wherein it is supposedly shown that multiple different qualitative states may underlie the same functional description. The second type is the type that Block described, where it is supposedly shown that a single qualitative state can be described by different functional descriptions.

[143] Ibid.
[144] Ibid.
[145] Lewis, 1999: 285
[146] Ibid., p. 288
[147] Ibid., p. 270

In order to reply to inversion arguments using the ability hypothesis, it is helpful to present the problem that spectrum inversion poses in terms of abilities. Suppose that the ability hypothesis is correct, and that we should see acquaintance with qualitative experiences as the ability to remember/imagine/recognize these experiences. This is compatible with intrasubjective spectrum inversion: we could recognize that the experience that we have when we look at the sky now has different qualitative properties as the one that we had before our intrasubjective spectrum inversion.[148]

The problem with intrasubjective spectrum inversion is not that it directly clashes with the ability hypothesis, it is that it goes against deep intuitions about our acquaintance with many types of qualitative experiences. For suppose that we should indeed account for the possibility that we may someday wake up and see the world around us in completely different colors. If this were so, our ability to make claims about future experiences would be limited: while the qualitative color-experiences that we have had while looking at lemons may have been very similar throughout our life up until now, we may have a whole other qualitative color-experience when we look at a lemon tomorrow. This means that the possibility of intrasubjective spectrum inversion may not only lead to skepticism about other minds, but also about our own mind: if we must account for the possibility of intrasubjective spectrum inversion, our previous and current color-experiences cannot help us predict what kind of qualitative color-experiences we will have when looking at familiar objects in the future. But this goes against our intuition that we *are* able to make those predictions, and past successes give us reason to think that we do indeed have that ability.

The unintuitive nature of intrasubjective spectrum inversion becomes more clear when we reflect on the link between qualitative aspects of experiences and other aspects of experience. When we reflect on some qualitative experience, it seems intuitive to think that this qualitative experience is somehow linked to a certain (or several) more objectively defined event(s). Even if the pain that we expect to experience during a visit

[148] A practical issue would be that our ability to recognize certain qualitative properties could not easily be described: if intrasubjective spectrum inversion is rife, we can no longer effectively use expressions like "being able to recognize the qualitative color-content of blue". This is because "the qualitative color-aspect of blue" is only a useful expression if we can assume that the intentional content of the color correlates with the qualitative content: we could only effectively use the expression "the color blue" to refer to the qualitative content if this content were always the same.

to the dentist is something that cannot be properly described in words, we seem to be quite confident in our ability to determine when that type of qualitative experience will occur. Perhaps the pain will be worse or less bad than we expected, but we are often able to predict what general type of qualitative experience comes along with some event. And we are also quite confident in our ability to predict whether other experiences will be painful, even if they are different from experiences that we have had up until now: I have never broken a finger, yet I am confident about the fact that this would probably be an at least somewhat painful experience. We can see that our abilities to predict qualitative experiences play a crucial role in our behavior and expectations, and the ability hypothesis fits well with that intuition. The idea of potential spectrum inversion goes against our intuitions regarding these abilities.

If we think more about this link between subjective experiences and objective events, intrasubjective spectrum inversion becomes even more unlikely. Before I turn to qualitative color-experience, I will talk about pain-experience. When we talk about pain-experiences, it is very natural to assume that the feeling of pain affects our behavior in various ways. Imagine that a neurologist invents a machine that can stimulate our nervous system, in such a way that it will be in the same state as during the experience of breaking your finger. It is only natural that we would dislike the idea of this machine being used on us, even if the resulting experience would not be accompanied by actual physical harm. And our choice to avoid the nerve-stimulation experiment does not only suggest that qualitative experience is a relevant factor in our choices, it also shows that we assume we are able to predict what type of qualitative experience the experiment will cause: we assume that that experience will be similar to the unpleasant one we have during experiences that involve pain.

Of course, a subjective report on one's behavior does not imply that the qualitative experience of pain is indeed what causes us to avoid pain: perhaps a pain-experience is something that merely happens to correlate with the type of events that can potentially harm a person, and it could be the case that the person is incorrect in their assessment that the predicted negative qualitative experience is a causal factor in their choice to avoid the neurologist's machine. But if this is the case, our intuition tells us that there must still be a correlation between the actual causal factor and the qualitative sensation. The existence of this correlation is a requirement for our ability to

successfully predict what type of qualitative sensation we will have under certain circumstances. If we did not assume that this correlation exists, our aversion to the neurologist's machine would not have a clear explanation. And if the link did not exist, it would not be clear why we would assume that it exists: if there was no link between certain events and the experience of pain, there would be no reason why we associate the experience of pain with certain events (like breaking a finger).

Even though it seems that intrasubjective spectrum inversion goes against our intuitions regarding our ability to predict future qualitative experiences, a defender could find a place for it. We could argue that our intuitive assumption that we can successfully reflect on/predict things about qualitative experiences can still be justified, if we hold that intrasubjective spectrum inversion is conceivable but highly unlikely. Suppose that scientists determine that a person's chance of undergoing sudden or gradual spectrum inversion at some point or during some period in their life is one in a billion: it would still be reasonable to expect that your qualitative color-experiences tomorrow will be similar to those today. The vast majority of people would be unaffected by it, and the potential of qualitative inversions would not affect their ability to predict tomorrow's color-experiences.

Now suppose that there are some unfortunate people for who the qualitative contents during certain experiences change very frequently: every day when they wake up, they find that the world around them seems to have different colors. When we wake them up and present them with a piece of colored paper (without any indications of what color it could be), they cannot give a convincing answer when we ask them what the color of the paper is called. They may know that the qualitative content of their current experience is similar to that of an experience that they had earlier, but their current experience lacks the representational content that would enable them to tell the color of the paper. Their qualitative experiences lack the type of consistency that is required for the ability of linking qualitative content to representational content. We would therefore be able to determine whether someone has undergone (or often undergoes) spectrum inversion by empirical tests such as the one with the piece of paper.

One could perhaps think that a combination of lacking abilities could mislead us into thinking that the person is capable of telling us the name of the color of the paper. If a test-subject's memory of color-perceptions shifts in harmony with their perception-shifts, they may never notice anything weird going on. Their memory-shifts prevent the ability to successfully see the similarity between past/current/future color-experiences, while their lack of ability to have consistent color-perceptions prevents them from determining representational content (such as the names of colors). The combination of the lacking abilities goes unnoticed by both the test-person and by those who do the empirical testing.

Does the scenario of lacking both consistent memories and consistent qualitative experiences actually make sense? It is not clear at all what exactly it is that supposedly changes, and how it affects their mind in any significant way. The internal aspects of the mind do not seem to differ in any significant way between the person whose memory and perception change in harmony and the person for who this change never occurs. They themselves will not notice anything, and their interaction with the external world does not change in any significant way: they are perfectly capable of pointing out the names of the colors of things around them. Since there seem to be no significant differences between the person with consistent memory and perception and the person with these supposed harmonized shifts, we may wonder whether the idea of a person with harmonized shifts even makes sense. Perhaps the idea that these shifting properties exist is a result of faulty theorizing: the assumption that they exist leads us to a problematic scenario, while rejection of the existence of these properties allows for a less complicated theory. And even if these properties *do* exist, they are not useful for those who wish to use inversion arguments: since the properties are not significant for the mind, the inability of a theory of mind to account for them is not a weakness of that theory. If the properties are not relevant for a description of the mind, a complete description of the mind can do without a description of these properties.

Involving the ability hypothesis also enables us to show the implausibility of interpersonal spectrum inversion. If we hold that people have some kind of access to the qualitative content of their color-experience, and that there is a link between qualitative experience and behavior, then a significant intersubjective difference in qualitative color-experiences (during a specific event) would manifest itself in empirically testable

ways. This fits well with the characterization of qualitative experience (and its role in our behavior and perspective on the world) that the ability hypothesis gives us. Instead of empirically testing a single person's relation to some color (like in the test with the colored paper), we could test whether multiple people have similar relations to some color. If we see that their behavior points toward similar abilities regarding the color, we could conclude that their color-experiences are similar. If we see significant differences in ability-related behavior (e.g. the ability to say the name of the color), we may be able to conclude that their color-experiences are different.

Determining whether two subjects have similar/different color-experiences is not easy: there are various obstacles for successful empirical testing. Perhaps two subjects have always had similar color-experiences, but one of them has memory-problems that cause him to think that his color-experiences change all the time. But even though the person may lack the ability to remember or predict experiences that are similar to his current one, we can still see the effects that it has on him at the moment. These effects may be subtle in the example of color, but more clear if we use other examples: a person in pain will likely show behavior aimed at avoiding whatever causes the pain. Moreover, the person will likely agree that their experience was one of pain at the moment, even though their later memory shift may lead them to think that it was not. may Here we see that even though a person may lack abilities regarding a certain qualitative experience, it is still possible to characterize their experience in terms of its causes and effects.

The possibility to characterize experiences without involving abilities may seem like a problem for the ability hypothesis. But it is not problematic that an experience may be characterized in terms other than abilities, as the ability hypothesis only holds that the knowledge of experience is something that we *can* characterize in terms of abilities. That qualitative experiences sometimes occur without related abilities is not necessarily a problem, because abilities need not necessarily be involved. The ability hypothesis only relies on the intuition that qualitative experiences are things that we *can* have a grasp of through our abilities related to them, and that we often do so.

While accepting the ability hypothesis (and its related implications regarding our acquaintance with qualitative experiences) may show that undetectable spectrum inversion is highly implausible, it does not show that it is inconceivable. But while the

ability hypothesis may therefore not give us a definitive way to reject inversion arguments, it does help us in building a plausible theory of mind that can account for the problematic role of qualitative experience. Acceptance of the ability hypothesis shows that spectrum inversion is not only highly implausible, but that the concept of spectrum inversion itself is suspect. If our abilities to recognize/remember/imagine qualitative experiences have been relatively successful up to this point, we can wonder how it could be that we would suddenly lose these abilities without detection by ourselves or by others. Having to account for the possibility of undetectable spectrum inversion would also require us to account for the possibility that our intuitions regarding our grip on qualitative experiences may be wrong, and those who use inversion arguments against functionalism would have to accept that unattractive requirement.

The ability hypothesis and Inverted Earth

If we analyze Block's case of Inverted Earth with the ability hypothesis in mind, I think that there are three types of ability that are directly relevant. The first is the ability to recognize the qualitative content of your color-experiences. The second is the ability to correctly apply color-terms to the objects in your experience on the basis of color-experiences. The third ability is the ability to apply color-terms to objects in the external world.

If you are (unknowingly) abducted to Inverted Earth and receive the inversion lenses, your ability to remember/recognize/imagine the qualitative content of color-experiences is not affected. The qualitative content of your color-experience when looking at the sky on Inverted Earth is the same as on Earth: even though the color-information of the objects on Inverted Earth is different than that of the corresponding objects on Earth, the inversion lenses make it so that the external color-information that reaches your perceptual apparatus is identical to that on Earth. You are correct when you say that you recognize the color-experience that you have while looking at the sky, as it is identical to the color-experience that you would have when you looked at the sky on Earth.

The potential issue comes in when we try to verify whether someone is able to correctly label the colors of things that they perceive. It may seem as if the abducted person is not, since their correct usage of color-terms may seem to be related to

properties of external objects: since they used to use the term "blue" in expressions related to the Earth-sky, we may be inclined to say that their term "blue" refers to the color of objects with the color of the Earth-sky (and the Inverted-Earth sky is not such an object). But if we hold that we are in some way able to sense similarities between qualitative experiences (which seems uncontroversial), we could also argue that their ability to correctly use color-terms is not reliant on the external world but on the link between their qualitative experience and the term that they apply to it. I will now explain this view further.

When we assess someone's ability to correctly use a color-term, we could first look at the link between some color-experience and the term that they use to refer to it. If a subject consistently uses the three terms "x", "y", "z" to respectively refer to the color of objects that give them color-experiences *XQ*, *YQ* and *ZQ*, then we can say that their usage of these terms refers to their corresponding color-experiences. Once we have established what terms they use, we can determine their ability to name colors by determining whether they use the names "x", "y", "z" when they are presented with objects that respectively give them color-experiences *XQ*, *YQ*, and ZQ.

If we assess a subject's ability to name colors with the proposed method, their ability to name colors on Inverted Earth will be the same as it was on Earth. If they say that "the sky is blue" while on Inverted Earth, they are correct: if "the sky is blue" used to mean that the sky was such that looking at it gave them a color-experience of blue, and the sky on Inverted Earth has the same effect (which is true, because of the inversion lenses), then they are right in calling the sky blue. We must remember that when they say "the sky" on Inverted Earth, we could say that this refers to [Inverted Earth-sky as seen through inversion lenses]. And since their perception of [Inverted Earth-sky as seen through inversion lenses] is identical to their perception of [Earth-sky without inversion lenses], they are correct when they use the same color-term for both. Under this interpretation of correct color-term usage, bringing a person to Inverted Earth does not change the functional buildup of their ability to name colors of objects as they perceive them.

There may seem to be something unnatural about my proposed method. It seems that when we say something like "the sky is blue", we are not merely talking about a property of our perception of the sky: a big part of it seems to be a claim about

properties of the actual sky in the external world. I agree that this is a big part of the statement, but I think that we could characterize description of this aspect as a separate ability: the ability to correctly refer to external objects. This way we can still say that people keep their ability to name colors during the whole process of [correctly naming colors on Normal Earth]->[incorrectly naming colors on Inverted Earth]->[correctly naming colors on Inverted Earth], while we also agree that there is a functional change regarding the relation between the subject and the external world.

I agree that the specific ability to correctly name things in the external world is where attempts at functional description may encounter a problem: while I have argued that the subject's ability to name the colors of perceived objects remains intact during the Inverted Earth-scenario, there does seem to be a change in their ability to assess properties of the external world. But the question is whether the inability to correctly name colors in the external world is something that is very problematic for a theory of mind. This question can be put as follows: is a functional change in the ability to correctly name colors of objects in the external world problematic for a description of the relevant aspects of a subject's mental states? I think that the answer is no. The main reason for this is that neither the behavior nor the qualitative aspects of the test-person's experience are affected. their mental state is not affected in any way that is relevant for the test-person themselves: they have no idea what they have undergone, and their behavior has not changed in any way. And during the whole process of [correctly naming colors on Normal Earth]->[incorrectly naming colors on Inverted Earth]->[correctly naming colors on Inverted Earth], neither their ability to recognize similar experiences nor the ability to name the colors in those experiences was affected by the move to Inverted Earth.

The conclusion of this part of the paper is that if we analyze qualitative experience in terms of abilities, inversion arguments do not pose a significant problem to analytic functionalism. The analysis of intrasubjective spectrum inversion shows that a sudden shift in color-experiences does not directly affect the ability to recognize similar experiences, and that any difficulties that would arise from this shift (e.g. lack of ability to correctly name colors) would be detectable in behavior. This also holds for intersubjective spectrum inversion, if we use empirical tests that can show the differences in the abilities of different people. Analytic functionalists who accept the

ability hypothesis are free to acknowledge that the Inverted Earth-scenario shows that some properties of behavior (e.g. transition in the correctness of using color-terms to refer to external objects) are not properly accounted for, because they can say that these properties are not relevant for a complete theory of mind: if we characterize qualitative experience in terms of abilities, a change in color-terms that are applied to external objects is not directly relevant. If the ability hypothesis works, we have a strong tool against inversion arguments. It would be valuable to have further independent arguments for the plausibility of the ability hypothesis, and therefore I will present some of these in the next part.

Part IV

The Plausibility of the Ability Hypothesis

In this part, I will attempt to show why a combination of analytic functionalism and the ability hypothesis can work as the basis for a plausible and complete theory of mind. This part consists of two smaller parts. In the first smaller part, I will reflect on the problematic intuitions regarding qualitative experience that underlie many arguments against analytic functionalism. I will then argue that AF+AH works with a characterization of qualitative experience that is less problematic. In the second smaller part, I will attempt to give a plausible explanation for the fact that qualitative experience seems so hard to characterize in a theory. I will use texts by Shoemaker, Quine and Dennett to form a potential picture of why it is that we have qualitative experiences, and why they may sometimes seem to evade successful formal description. I will also show how AF+AH fits well with this picture, and that even though analytic functionalism may seem unintuitive in some parts, this is not a reason for rejecting it.

Problems in the traditional view on qualia

In 'Quining Qualia'(1993), Dennett wants to show that the traditional notion of qualia that is often appealed to is fundamentally flawed, and that an attempt to improve this notion disarms many arguments that rely on it.[149] He presents and attacks fifteen intuitions that often underlie the problematic notion of qualia, and I will now reflect on some of his strongest points.

The first intuition that he focuses on is one that I have already covered earlier: the intuition that intersubjective and intrasubjective shifts in qualitative experience are possible.[150] Dennett says that the big mistake in this intuition is the thought that qualia can be isolated from everything else in experience.[151] But the fault in this idea is not only the thought that we are capable of performing this isolation: the fundamental mistake is in the thought that there would even be such an isolated qualitative property at all.[152] Dennett thinks that there are no such properties: we cannot isolate the qualitative smell/taste/sound that an individual experiences from the experience, as there is no such isolated thing.[153]

Another often-held intuition regarding qualia is that we have infallible access to them: it may seem as if having some qualitative experience gives us authoritative access to properties of that qualitative experience. To show that we do not have infallible access to our supposed qualitative experiences, Dennett introduces an example of two experienced coffee tasters whose job at a coffee factory is to ensure that the taste of the coffee remains constant. One of them (Chase) says that he has lost his appreciation for the taste of the coffee: he thinks that the taste of the coffee has remained constant over the years, but that his own preference for tastes has changed.[154] The other taster (Sanborn) is in a different situation: he has also lost his appreciation for the coffee, but he thinks it is because his taste-buds or some other part in his perceptual machinery has somehow changed. He is convinced that he still has the same preference for tastes, but

[149] Dennett, 1993: 381
[150] Ibid., p. 383
[151] Ibid.
[152] Ibid., p. 384
[153] Ibid.
[154] Ibid., p. 388

that his sense of taste has changed so that he can no longer experience them like he did before.[155]

We could accept that Chase and Sanborn are infallible about their qualitative experience, and that they are both right in their assessments. But there are alternative explanations for the changes in their experiences:[156] Chase might be mistaken about his change in taste preference, and his qualia may have slowly shifted over the years. And Sanborn's standard may have shifted over the years: he thinks that his earlier experiences with the coffee were better, because nostalgia has tinted his memory. It is also possible that both of them are partly right in their assessments, while they are also partly affected by the alternatives.[157] It seems that these alternative options could just as well be what is going on, and we have no clear reason to think that Chase and Sanborn are right in their assessments.

In order to make a reliable statement about the properties of our experience, Dennett thinks that we should regard our self-reports as judgments: "a subject's experience has the quale *F* if and only if the subject judges his experience to have quale *F*."[158] The quale is then introduced through the judgment of the subject.[159] If we regard self-reports regarding qualitative reports as judgments, qualia should be regarded as mere theoretical constructs.[160] We could think of empirical tests that would show whether Chase and Sanborn are right in their assessments (such as blind tastings).[161] The results of these test could support or oppose their claims, and give us some insight in whether they indeed have infallible access to their qualitative experience.[162]

There are complications for the evaluation of claims regarding qualia. We can imagine a surgery where the connections of a person's taste buds are shifted, so that things now taste different for that person.[163] It is possible that other parts of the person's perceptual system that come *before* the judgment regarding qualia already

[155] Ibid., p. 390
[156] Ibid., p. 391
[157] Ibid.
[158] Ibid., p. 392
[159] Ibid.
[160] Ibid.
[161] Ibid.
[162] Ibid., p. 393
[163] Ibid., p. 394

compensate for the shift in the taste buds.[164] This would mean that the qualia remain the same, as the physiological facts still result in the same input for the part of the process where qualia are introduced. But it could also be the case that compensation takes place *after* the part where the qualia is introduced.[165] The person may think that lemons taste as sour as ever (even though his taste buds now produce output that would previously have been processed as sweet), because his memories and language regarding the taste of lemons have compensated for the new taste. In the latter case the supposed quale involved in tasting sour things has changed, even though the person thinks that it is still the same.[166]

The two possible cases presented above show that attempts to evaluate claims about qualia based on empirical methods can be complicated. The subject himself cannot settle which of the two options is applicable to his situation, as he has no faculties to determine whether it are his memories or his qualia that are different.[167]A scientist can come to both conclusions, as their conclusion is determined by how they have chosen to exactly define which properties of the cognitive process are qualia.[168] The scientist could choose to say that the qualia have changed after the surgery (and that the subjects conviction that they are still the same must be the result of memory-compensation), or that compensation has taken place before the step in the process where the qualia are judged (meaning that the subject is right in his judgment that his qualia have remained the same).[169] Due to the fact that the subject cannot justify his conviction that one of these is the right answer, the scientist is free to choose their own conception of qualia that fits one of these conclusions.[170]

Dennett sees a problem in the traditional idea that experiences give us access to some sort of intrinsic property of something (like an intrinsic taste or smell). He gives an example to show that even when things may seem intrinsic, careful analysis can often show that they are relational.[171] An experienced beer drinker may say that beer is an acquired taste: most people do not enjoy their first sip of beer, and if every subsequent

[164] Ibid.
[165] Ibid.
[166] Ibid.
[167] Ibid., p. 395
[168] Ibid.
[169] Ibid.
[170] Ibid.
[171] Ibid., p. 397

sip would taste similar to the first one, no one would keep drinking beer.[172] This statement seems plausible, and it presents a big problem for defenders of traditional qualia. The taste of the first sip may have seemed to produce acquaintance with some sort of 'intrinsic taste of beer', but the fact that one's appreciation for beer can change during later tastings means that the 'intrinsic qualities' of our beer-drinking experience are actually relational properties: the experience depends on both independent properties of the beer itself and on your taste-perception of these properties.[173]

The ability hypothesis versus Dennett

I think that Dennett's points are no problem for the ability hypothesis, and that we can accept the ability hypothesis while we also accept Dennett's perspective on qualitative experience. I will now reflect on three issues that Dennett's view may seem to present for the ability hypothesis, and I will explain why I think that they are not actually very problematic.

The first potential issue is Dennett´s claim that we cannot isolate some sort of qualitative aspect of experiences. We can accept both this claim and the ability hypothesis, as the ability hypothesis does not necessarily hold that there is some isolated qualitative aspect of experience. When we say that we are able to remember/recognize/imagine some experience, we do indeed suggest that we are acquainted with some sort of qualitative aspect. But we do not need to hold that this aspect is some fine-grained isolable thing, as our recognition of experiences may not require such a fine-grained access to some supposed quale. When we claim that we are acquainted with the experience of seeing blue, we do not have to hold that there is some special specific attribute of seeing-blue-experiences that we are acquainted with. We only need to have a somewhat clear idea of what we experience when we are confronted with blue, and under what circumstances this experiences would come about. We would also need to be able to in some way imagine what effect the involvement of blue-experience would have on our general experience at some moment. But we would not need to be acquainted with some specific type of isolable quality of blue-experiences that supposedly make these experiences into blue-experiences.

[172] Ibid.
[173] Ibid.

The second issue that Dennett addresses is our fallibility in our reflection on our qualitative experiences. Examples such as that of Chase and Sanborn may seem to suggest that our ability to recognize/imagine/remember experiences may not be reliable. But this problem can be softened by noting that the ability hypothesis does not directly make claims about our success in recognizing/imagining/remembering all types of experiences. Perhaps our ability to accurately recognize similar (or dissimilar) taste-experiences over time is simply not good enough, but that does not mean that we do not have the ability to roughly imagine/recognize/remember many other things.

My point becomes more clear when we look at the example of coffee-tasting. When we say that our current experience while drinking coffee (after drinking the exact same type of coffee for years) is different than it was years ago, we do not need to make any claims about why it is that it is different. We could simply say that we do not recognize the current experience as an experience that we have had before, regardless of whether we actually *did* have the experience earlier or not. It is irrelevant whether it is our taste-preference that changed, or our perceptual apparatus, or some combination of both. We may have the rough ability to recognize that it is the same coffee that we are drinking, while lacking the ability to imagine exactly what it was like when we drank the coffee years ago. And this is fine: upon reflection, many people will acknowledge that their acquaintance with the exact experience that they had while tasting coffee years ago is not very accurate or reliable.

A potential problem comes from the fact that we have to account for memories that are blatantly false. Suppose that you drank some type of coffee many years ago, and hated the taste. Now you drink it again, and love the taste. Your memory of the last time you drank it may be flawed, and you may think that your positive experience with the taste at the moment is just like the experience you had back then. How should we account for this type of fallibility? I think that analytic functionalism can be helpful here, as functional analysis of both experiences can show us why we are mistaken and why our perspective on the experiences is not infallible. Suppose that years ago, your behavior while you drank the coffee was carefully documented. You were asked various question to find out what you really thought about the coffee, and this (alongside with further behavioral analysis) was used to determine that your experience was very negative. When you are now confronted with the results of this analysis, you may be

shocked to find out that the situation was very different from how you remember it. But would you reject the analysis? It seems reasonable to say that many people would be prepared to reject their memory and acknowledge that their previous experience was not as they thought. And even if they would not do so, we could reject their claim about their memory ourselves.

The third relevant issue that Dennett notes is that we often do not become acquainted with intrinsic properties of things. This issue is not a problem for the ability hypothesis: our ability to recognize/remember/imagine does not rely on access to supposed intrinsic properties of things. When we say that a current experience is similar to an earlier experience, we simply say that something about the experiences (whatever it may exactly be) is similar enough that it warrants the claim that they are alike. Whether the properties that we access in our experiences are intrinsic or relational is not relevant, as long as we can recognize that experiences are similar or dissimilar.

It seems safe to conclude that the notion of qualitative experience used for the ability hypothesis is not affected by the problems that Dennett describes. While the ability hypothesis is an attempt to find a characterization for our acquaintance with qualitative experience, it does not involve the problematic notion of qualia that Dennett attacks. The claim that we are able to recognize/imagine/remember qualitative experiences does not say much about the exact character of these experiences, and it does not involve infallible access or commitment to acquaintance with intrinsic qualities. It only requires acceptance of the (relatively broad) idea that there are certain similarities between qualitative experiences, and that we can often be aware of these similarities in some way.

Some useful reflections on similarity

In order to further defend the plausibility of the ability hypothesis, I will first present some viewpoints regarding the concept of similarity of experiences. Similarity (and dissimilarity) play a fundamental role in evolutionary processes: an organism's chance to survive is dependent on its ability to behave in a way that suits its environment, and this requires the ability to recognize whether some object or process in its environment is similar to the objects or processes that are either positive (e.g. food, shelter) or negative (e.g. predators) for the organism. To show why the concepts of similarity and

ability play a role both on a relatively simple level (like in the case of insects that search for food) and on a more complex level (like in our conscious reflection on experience), I will now present the viewpoints of Quine (1969) and Shoemaker (1975) on the role of similarity. After that, I will reflect on Dennett's (2017) perspective on our ability to be aware of similarities. I think that a combination of these three views will give us a plausible characterization of our experience and (in)abilities (including our inability to always successfully convey experiences through description), that fits well with analytic functionalism and the ability hypothesis.

Awareness of similarities

In 'Phenomenal Similarity'(1975), Shoemaker attempts to provide an explanation of what qualitative similarities are and how we should analyze them. The idea that we have immediate intuitive awareness of the similarity and differences between our phenomenal states comes with a problem: how are we able to see similarity, and how can we know that some qualities are similar to others? Shoemaker refers to Quine's description of 'innate quality spaces' as a start for the solution of this problem.

> A response to a red circle, if it is rewarded, will be elicited again by a pink ellipse more readily than by a blue triangle; the red circle resembles the pink ellipse more than the blue triangle. Without some such prior spacings of qualities, we could never acquire a habit; all stimuli would be equally alike and equally different. These spacings of qualities, on the part of men and other animals, can be explored and mapped in the laboratory by experiments in conditioning and extinction. Needed as they are for all learning, these distinctive spacings cannot themselves all be learned; some must be innate. (Quine, 1969: 123)

While Shoemaker presumes that Quine would use behavioral terms to judge the accordance between a creature's innate quality spaces and relevant groupings in nature, Shoemaker himself will provide an account with mentalistic terms.[174] A proper innate spacing can then be described as follows: if a creature has the proper spacing for relations of similarity (or difference) between things, then that means that its perceptual

[174] Shoemaker, 1975: 15

experiences of similarity between things corresponds to actual similarities between those things.[175]

It may seem odd that the subjective quality spaces that we use to group similar things together is so harmonious with the relevant groupings in nature, but Shoemaker follows Quine in saying that this is the result of natural selection.[176] Creatures that lack the ability to make inductions that accord with the relevant groupings in nature are less likely to survive. They are unable to process their environment in a way that favors their survival, while there are other creatures that *are* able to do this processing. This means that natural selection favors the latter group of creatures, as they are better suited to live in their environment and have a better chance of creating offspring.[177]

The reason for our awareness of similarities

While Shoemaker reflects on our ability to be aware of similarities between perceptual experiences, he does not give a clear reason for why we are consciously aware of them. The explanation for awareness of similarities between sensory experiences seemed very plausible: any organism without this ability would go extinct. But this reason does not extend to awareness of perceptual similarity: we can safely say that many organisms do not have this type of awareness, yet are still able to survive as a species. This raises the question of why humans *do* have this ability, for which Dennett provides a possible answer.

One of the topics that Dennett addresses in *From Bacteria to Bach and Back* (2017) is the role of our 'sense of self'. Our sense of self seems to have developed for the process of communication: in order to successfully communicate relatively complicated things, we need to see ourselves as a subject just like how we see others.[178] Because our communication is more complex than the primitive forms of communication found in other animals, our sense of self is also more complex than that of other animals.[179] Not only do we have to keep track of which objects are part of our own body (just like simple organisms), we also have to monitor our awareness of certain pieces of information and

[175] Ibid.
[176] Ibid. p. 14
[177] Ibid.
[178] Dennett, 2017: 344
[179] Ibid.

process whether we want to communicate this information in any way.[180] Dennett defines our stream of consciousness as an 'edited digest' of activities in our brain, available because of its role in communicating our thoughts (in the broadest sense) to others and ourselves.[181]

Dennett's view on consciousness fits well as an answer to the question of why we are aware of the similarities between perceptions. While simple organisms only require the ability to determine sensory similarities in the relevant basic quality spaces (e.g. colors, smells), humans have evolved in a way that makes communication a very important aspect of life. For communication, the ability to recognize similar sensory experiences is not enough in itself: we also need to be able to process and interpret this similarity in a way that enables us to convey it to others or ourselves.

Dennett notes that our access to our thinking and the processes involved in it is not categorically different or better than our access to other bodily processes (such as our digestive system).[182] While awareness of similar experiences may only take place in more complex organisms, the awareness itself is not necessarily a more complex process. Dennett's notion of 'local competences' refers to relatively simple brain activities, while 'global comprehension' refers to more complex activities such as the ability to communicate things, or to plan ahead, or to imagine a variety of potential future situations.[183] If we follow Dennett in his view that consciousness is the result of a complex structure made out of local competences[184], we can say that awareness of similar experiences is the result of the structure, while other types of awareness (like the basic awareness of similar colors) have different roles.

It is possible to describe our awareness of similar experiences with the help of Quine's notion of quality spaces. While simple organisms only have a limited range of relevant quality spaces (e.g. colors or shapes), effective communication requires us to be able to determine which of our perceptions are relevant in communication. We could say that besides the sensory quality spaces, there are also quality spaces for communication. But instead of being related to the relevant basic groupings in nature, these quality

[180] Ibid., p. 344
[181] Ibid., p. 345
[182] Ibid., p. 346
[183] Ibid., p. 341
[184] Ibid., p. 345

spaces are related to the relevant social groupings. The ability of simple organisms to be aware of the right sensory quality spaces is needed for their survival in nature, while our ability to be aware of the right perceptual quality spaces is needed for a successful life in a social context. Since the ability to successfully communicate (whether it is to cooperate, deceive, or in some other context) gives an advantage over others that lack this ability, it is plausible that the process of natural selection has favored those with the best communicative abilities.

AF+AH in the context of Quine, Shoemaker and Dennett

While it seems to be possible to find a place for qualitative experience in an evolutionary account of the mental, the connection between qualitative experience and evolution is not directly clear. We must still explain why we have evolved to have some sort of rough conscious access to qualitative experiences (either through direct exposure or through introspection/remembering), and why at the same time there are aspects of these experiences that we cannot effectively describe.

When we see our cognitive abilities as the result of evolutionary processes, it becomes clear why it is hard to describe some aspects of experience. Since successful species adapt to their environment, it is to be expected that our higher cognitive functions are also fine-tuned in terms of value for our survival. While our abilities such as abstract thinking and communication may allow for reflections that are not directly beneficial to our survival (like reflection on qualitative experience), it is only to be expected that they do not always provide success in those cases. And I think that if we take a further look at our capabilities to reflect on things and communicate them, it will become clear that the limited success of our ability to communicate about experiences is based on the limited role of this communication in survival.

Consider the example of communication related to color. When we are confronted with a yellow object, this confrontation can affect us in all sorts of ways. We may recognize the color and think of other things that have the color. Through introspection, we may focus on the color itself. The experience may invoke all sorts of other thoughts and mental processes that affect us in many ways, both consciously and unconsciously. But when we communicate to others about color, we can only do so in functionally definable terms: “there is a yellow object behind this wall”, “yellow is the

color of bananas and the sun", "I like the color yellow", "an eye-condition prevents me from distinguishing yellow and orange", etcetera. We can only communicate our relation to the color yellow by saying things about the effect that the color has on us, and usually only in relatively vague terms.

Our inability to clearly communicate the complete functional role that experiences have for us also works the other way around: we cannot learn the full effect of an experience on our mental state by being told about it. If another person tells us about their thrilling experience of riding a rollercoaster, we can only understand them insofar as we are able to relate the tale to our own experiences. Many intense experiences affect us in various ways, and it is clearly wrong to assume that this whole process could be communicated. But it would be a mistake to use this incommunicability as an argument for the existence of some type of 'phenomenal information', as we have seen earlier that this notion is very problematic in itself. There seems to be no clear advantage in having the ability to communicate *what it is like* to see some color, so it is not weird that we have not evolved to be able to do so. And while this inability to communicate some aspect of experiences may lead to a type of mystery regarding qualitative experience, it seems wrong to take this mystery as a reason for accepting that there must be some kind of phenomenal aspect to the world.

We are now at a point where we can look at inversion arguments from a functionalist perspective again. If we look at color-perception from a functional perspective, undetectable color-inversion is implausible. If our conscious experience is based on all the local parts of our brain working together to effectively live in our environment, the introduction of some phenomenal extra layer on top of our functional buildup is unnecessary: our ability to reflect and report on experiences is all that we need to account for, and the limits of these abilities are no reason to introduce phenomenal qualities.

It is, of course, not the case that our mind should necessarily have the most simple structure. It is conceivable that there are qualitative phenomena that somehow occur alongside our physical/functional processes (like the nomological danglers that Smart spoke of). Jackson (1982) argues that even though qualia may not be conducive

to survival, they *can* be a byproduct of traits that are conducive to survival.[185] But while an appeal to parsimony may not be the ideal reason for accepting functionalism, it may be possible to show just how very unlikely it would be that there is some phenomenal layer of experience.

In a certain sense, sensory information 'seeps' through our cognitive systems: if we were to confront a test subject in a controlled environment with a blue mug, the brain-processes that start would be considerably different from the processes that would have started from confrontation with a yellow mug. This may not directly show in external behavior, and it is up to scientists to show the extent of the differences in the brain-processes. But it is clear that slight variations in the external environment can lead to very complex differences between inner processes, and it would be bizarre to think that our personal experience gives us better authority to speak about these differences.

Our brain is so complex that we cannot easily, if at all, grasp how it works. We may be able to create complete theories to describe the brain, but a single person will not be able to fully comprehend what is going on in their brain at a certain time. If we describe the brain through scientific theories, we can find all sorts of intricate functions and structures. We could find how various degrees of pain affect behavior, and how confrontations with different pieces of color-information start different processes in the brain. It is clear that the resultant body of knowledge will be vast and complex, and it may sometimes lead to descriptions of experience that do not seem to fit with some intuitions regarding experience. But instead of thinking that there is some phenomenal aspect to the world that cannot be grasped in functionalist terms, it seems more appropriate to think that our functional buildup is so complex that we cannot fully grasp how it leads to qualitative experience.

[185] Jackson, 1982: 134

Conclusion

The goal of this paper was to defend the thesis that *analytic functionalism provides an appropriate framework for a plausible and complete analysis of mental states*. To do this, I have tried to show that a functionalist approach is the most plausible approach to the analysis of mental states. Many philosophers would agree that some degree of functional analysis is useful in the description of mental states, but most discussion revolves around the question whether a functional analysis can provide a *complete* description of mental states. I agree that there are many intuitive reasons to doubt this, but that thorough reflection on the way that we approach mental states shows that functional analysis can provide a complete description of mental states.

A definitive proof of Lewis's view that mental states are necessarily identical with physical states would be the most convincing argument for this view, but I have not been able to create such a proof. Instead, I have argued in favor of analytic functionalism in three steps. The first step was to show how accepting the ability hypothesis would allow us to see the implausibility of inversion arguments, thereby refuting a main argument against analytic functionalism. The second step was showing that the characterization of qualitative experiences used in the ability hypothesis is able to withstand many arguments against the traditional conception of qualia. The third step was showing that the combination of AF+AH fits well in a evolutionary account of the mind, and that it is to be expected that some aspects of qualitative experience cannot be fully grasped based on theoretical descriptions.

I think that the combination of analytic functionalism and the ability hypothesis gives us a strong basis to argue against the plausibility of inversion arguments. Analytic functionalism by itself gives us an effective way to characterize many parts of the mind, albeit in a way that may leave open some doubts about qualitative properties. By introducing the ability hypothesis, it becomes possible to bridge the gap between the purely functional and the qualitative aspects of experience. The ability hypothesis gives us a way to characterize the qualitative aspects of experiences in a way that fits well with many of our intuitions, without falling prey to the problematic notion of qualia that is required for classic inversion arguments. Block's example of Inverted Earth may show that there is something related to the mind that is not captured by AF+AH (potential

transitions in the correct reference to the colors of things in the external world), but I think that this weakness is not a significant problem for a theory of mind. I think it is safe to conclude that the combination of AF and AH is plausible enough that defenders of inversion arguments should doubt whether they are prepared to uphold the intuitions that underlie their view.

When we consider the problems in alternative positions and the fact that functionalism can be defended against many of the main doubts, I think that we can conclude that analytic functionalism can indeed provide a plausible and reliable analysis of mental states. But I think that there is another strength of functional analysis that deserves a brief mention. It is clear that there is a significant gap between the folk-psychological way of understanding the mind on one side, and the scientific way on the other. While the last 100 years are filled with scientific developments regarding the brain, these developments do not always translate well into folk-psychological notions. In order to bridge the gap, it is helpful to have a theory that puts folk-psychological concepts into a scientific perspective. I think that analytic functionalism is a useful theory in this regard, as it allows for a practical theory of mind that is supported by science.

One could argue that the folk-psychological influences in analytic philosophy are a weakness, and that our common-sense notions about the mind should be rejected in favor of a purely scientific theory. There is some logic behind the eliminativist view: since our common-sense view on the mind is bound to contain some flaws, a purely scientific view would surely be more reliable. But the complete rejection of common-sense notions is problematic: if the scientific theory is too far removed from our practical understanding of our mind, we can doubt whether it is suitable for answering the questions about the mind that we are interested in. Scientific findings can help to correct mistakes that we make in our everyday assessments regarding our own mind and the minds of others, and we should always pursue a fuller understanding of the mind by scientific means. But we have to account for our inability to effectively use a complex and unintuitive scientific theory. A theory that bridges the gap between science and common sense is a valuable compromise, and I think that analytic functionalism (in combination with the ability hypothesis) is suitable for this purpose.

While analytic functionalism is useful, it is not perfect. I think that I have found a reasonable explanation for the supposed elusiveness of qualitative properties, but this may not be convincing to everyone. I think that my explanation is plausible and not as problematic as alternatives, but there is still room for doubt. A possible reason for this doubt could be that our supposed access to qualitative properties seems very natural and intuitive, while rejection of this supposed access involves a lot of reasoning. But while this may be the case, I think that it is still possible to convince others of the plausibility of analytic functionalism. It is unlikely that a single paper will lead to sudden rejection of deeply-rooted intuitions regarding the mind, but it may be possible to stir some doubts in some people. And if this paper helps anyone to become more sympathetic towards analytic functionalism, then it will have served a valuable purpose.

Bibliography

Armstrong, D. M. (1968). *A Materialist Theory of the Mind*, Londen: RKP

-(1980). The Nature of Mind. In Block 1980a: 191-199

Block, N. (1980a). *Readings in the Philosophy of Psychology, Volume I* (ed.). Harvard University Press

-(1980b). What is Functionalism? In Block 1980a: 171-184.

-(1980c). Troubles With Functionalism. In Block 1980a: 268-305

-(1990). Inverted Earth. *Philosophical Perspectives* Vol. 4: 53-79

-(1996). Mental Paint and Mental Latex. *Philosophical Issues* Vol. 7: 19-49

Broad, C. (1925). *The Mind and Its Place in Nature*, London: Kegan Paul

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory, New York: Oxford University Press*

-(2010). *The Character of Consciousness*, Oxford University Press

Churchland, P. (1986). *Neurophilosophy*: *Toward a Unified Science of the Mind-Brain.* MIT Press

-(2002). *Brain-wise: studies in neurophilosophy*. Cambridge, Mass. : MIT Press

Dennett, D. (1991). *Consciousness explained*, Boston: Little, Brown

-(1993). Quining Qualia. In *Readings in Philosophy and Cognitive Science*, Goldman, A. (ed.): 381-414. Massachusetts: MIT Press

-(2017). *From Bacteria to Bach and Back: The Evolution of Minds*, Penguin Books

Feigl, H. (1967). *The Mental and the Physical: The Essay and a Postscript*, Minneapolis: University of Minnesota Press

Frege, G. (1956). The Thought: A Logical Inquiry. *Mind*, LXV, 259 (July 1956): 289-311

Harman, G. (1990). The Intrinsic Quality of Experience. *Philosophical Perspectives,* Vol. 4, Action Theory and Philosophy of Mind: 31-52

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127- 136.

Kripke, S. (1980). Excerpt from "Identity and Necessity". In Block 1980a: 144-147.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press

Lewis, D. (1966). An Argument for the Identity Theory. Reprinted in *The Journal of Philosophy* Vol. 63: 17-25 (1999)

-(1980). Mad Pain and Martian Pain. In Block 1980a: 216-222

-(1999). What Experience Teaches. In *Papers in Metaphysics and Epistemology Volume 2*: 262-290. New Jersey: Princeton University Press

Locke, D. (1968). *Myself and Others: A study in Our Knowledge of Minds*. Oxford University Press

McFarland, D. (1989). *Problems of Animal Behaviour*. Harlow, Essex, UK: Long-man Scientific and Technical

Nagel, T. (1974). What Is It Like To Be a Bat? In *Mortal Questions* (2012): 165-180, Cambridge: Cambridge University Press

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In *Mind and Cognition: A Reader*, Lycan, W. (ed.)(1990): 490-499, Oxford: Blackwell

Putnam, H. (1980). Brains and Behavior. In Block 1980a: 24-36

Quine, W. (1969). *Ontological Relativity and Other Essays,* New York: Columbia University Press

Ryle, G. (1949). *The Concept of Mind*, Londen: Hutchinson. Page references are to the 2009 60th-anniversary edition, London: Routledge

Schlick, M. (1959). Positivism and Realism. In *Logical Positivism*, A.J. Ayer (ed.): 82-107, New York: Free Press

Shoemaker, S. (1975). Phenomenal Similarity. *Crítica*. Vol.7(20): 3-34

-(1982). The Inverted Spectrum. *The Journal of Philosophy*, Vol. 79: 357-381

Smart, J.J.C. (1959). Sensations and Brain Processes. *Philosophical Review* 68: 141-156

Tye, M. (1994). Qualia, Content, and the Inverted Spectrum. *Noûs*, Vol. 28, No. 2: 159-183

Watson, J. B. (1913). Psychology as the Behaviorist Views it. *Psychological Review* 20: 158-177

Wittgenstein, L. (1968). Notes for Lectures on 'Private Experience' and 'Sense-data', *Philosophical Review* 77: 275-320