Consciousness, explainable?

Jurriaan Schuring

Student number: 519 008
Word count: 9131
Study: Double Degree Philosophy
Faculty and university: ESPhil, EUR
Main study: Nanobiology, TU Delft & Erasmus MC
Date: 26-06-2022
Supervisor: Prof.dr. F.A. Muller
Advisor: Dr. T.K.A.M. de Mey

Contents

1. How can consciousness be explained? 2
2. The hard problem of consciousness 4
3. Supervenience and reduction 6
Logical and natural supervenience 6
Logical supervenience problems 7
Logical Supervenience of the Macroworld 8
Reduction and consciousness 9
4. Zombies and non-physical properties 10
Chalmers' view 10
Circularity? 10
Meaning of consciousness 11
Conceivability 11
A sceptical problem 12
Dennett's view 12
5. Computationalism 14
The causality of information states 14
A Chinese thought experiment 15
6. Consciousness beyond experience 16
7. Conclusion 17
Works Cited 18

1. How can consciousness be explained?

The nature of our mental lives seems to have puzzled and intrigued many people throughout history. Consciousness especially has been something baffling to philosophers and scientists alike, causing a wide abundance of theories trying to explain our conscious experience. A theory of consciousness is not only interesting because it would allow us to reach a deeper understanding of our daily experience. Additionally, it is beneficial for pathological and moral issues. Take mental illnesses, people classified as bipolar or schizophrenic are reported to have "altered states of consciousness" in which they can show abnormal behaviour (Bachelor and Roussel 2001). Exactly what causes these states remains unclear and possible treatments are limited. A theory of consciousness could be used to elucidate potential causes and thereby also possible therapies. Additionally, it might help in predicting cases, which can benefit people with a high risk for mental diseases. For example, someone with a high risk for schizophrenia could be prepared for a first episode or it could be even prevented, potentially decreasing many harmful effects (McFarlane 2011).

A moral aspect for which a theory of consciousness is helpful, is understanding to what degree/in what way animals are conscious. That animals are conscious in one way or another seems to be widely accepted (Allen and Bekoff 2017). However, to what extend animals have consciousness, whether it is anywhere remotely comparable to human consciousness, still needs further inquiry. One potential "real world" impact of a theory of consciousness could be on the bio-industry and animal rights. If consciousness of some animals appears to be more human-like than expected, then certain practices in bio-industry would perhaps become forbidden. In the future we might have to deal with similar issues for robots or computer programs, or for both. Will Alternative Intelligence (AI) make computers or robots conscious? Is that even possible? If these questions can be answered affirmatively, then a moral foundation needs to be developed to prepare for ethical issues related to conscious machines. Would it be ethical to turn them off? Is slavery of robots allowed? A theory of consciousness can show whether any of this is actually intelligible or just a fantasy for television.

The possible applications described above are however still futuristic, as a well established, widely accepted theory of consciousness has yet to emerge (Gulick 2021), despite much progress that has been made in the fields of neuroscience, AI and psychology towards a more profound understanding of biological functions of the brain. This is not only due to insufficient empirical knowledge of the supposed locus of consciousness, the brain (so far as you can speak of a "locus" of consciousness), but also due to conceptual disagreements. First of all, definitions of consciousness vary widely (Gulick 2021), so not everyone agrees on what it is that needs to be explained. Secondly and related to the previous issue, the necessary mode of explanation is not agreed upon by philosophers, with a host of different contenders.

One might argue that these are just philosophical issues for philosophers to squabble about and for them to fill bookshelves with their arguments. In the end it will be neuroscientists (of course) who will show the intricate workings of the brain, the grand seat of consciousness. In due time, the neural pathways of the brain will be mapped, modelled and analysed, from where it will be shown what consciousness is, how it works and whether we can replicate it in robots. Then finally neuroscience will even solve all philosophical issues, ranging from morality to epistemology as some have argued (Hacker and Bennett 2022, 458).

However, this is misconceived. A theory of consciousness will need to explain conscious experiences which can be described with psychological predicates. A theory can be seen as a logical framework connecting concepts. Whether or not these logical connections make sense is a task for logic and philosophy to find out. Empirical science does not establish this logic itself, rather it presupposes it so

it can then use it to test theories. If empirical science tries to establish truth or falsity about a senseless or contradictory statement, then it will not achieve anything (Hacker and Bennett 2022, 467). Just like: “The king of France is bald”, is not an interesting statement to test empirically.

Philosophy is also needed the part in which a theory is tested. How do we know a certain statement is true? This is what epistemology tries to answer. Furthermore, how do we know we have explained consciousness? What do we mean by explaining consciousness? Answers to these questions are useful for establishing a theory of consciousness, but they cannot be answered by empirical science. Once again because epistemology sets up logical connections between concepts (Hacker and Bennett 2022, 466), which is part of philosophy. Hence, philosophy is needed next to neuroscience to work towards a theory of consciousness. Both to provide the conceptual clarity and the proper epistemology. This will prevent ending up at a theory similar to Descartes’ substance-dualism.

In fact, many neuroscientists seem to be stuck in the 17th century philosophy of Descartes or Locke. For example, Kandel writes about the so-called binding problem, he asks: “How is information carried by separate pathways brought together into a coherent visual image?” (Kandel, Schwartz and Jessel 2012, 492). The idea that separate pathways need to somehow unite their information in order to create a visual image to “show” to our consciousness is very much reminiscent of Descartes’ mind: the Cartesian theatre (Dennet 1991, 107). Surely Kandel is not proposing a substance dualism similar to Descartes. However, the form in which he puts the problem is very much Cartesian, bringing along all its logical inconsistencies (Hacker and Bennett 2022, 148). Other neuroscientists adhere to the view that colour is essentially something created by the brain, which is just a continuation upon Locke’s secondary qualities (Hacker and Bennett 2022, 466). To avoid mistakes such as these, conceptual clarity of the theory is essential.

In this thesis, I will investigate possible explanations of consciousness, focusing on perceptual consciousness. There are of course also other aspects to consciousness. However, it is beyond the scope of this thesis to discuss all aspects of consciousness and their possible explanations. I will take a closer look at the two main metaphysical theories of consciousness, dualism and physicalism (Gulick 2021). For dualism, I will investigate Chalmers and his property dualism. For physicalism, I will take a look at computationalism. (Interestingly Chalmers also suggests to use computational like explanations, but without it serving as the ontological basis (Chalmers 1996, 276-308)). In the first chapters, I will critically investigate Chalmers’ reasons for dualism. Furthermore, I will also discuss a reductionistic approach to explain consciousness. Later, I will focus on the approach of computationalism and point out its deficiencies. Finally, I will argue for a non-reductionistic approach. I believe there cannot be an ultimate physical explanation of consciousness. This approach should focus on physical conditions in the brain for there to be consciousness, without jumping ahead to assertions about causes and the ontology of consciousness.

2. The hard problem of consciousness

One of the key features of consciousness is its subjectiveness. All perceptions and sensations we experience, have some sort of subjective aspect to it. Another way to put it is that, as Chalmers writes, there is a qualitative aspect to experience (Chalmers 1996, 4). There is a certain way an apple tastes and feels when you bite it. Music can be awe-some or the colours of a painting can be extremely vivid. More day-to-day examples are smells coming from the kitchen (or the toilet) or the sight from your window. In all these experiences there seems to be something personal which might be different from person to person. This is the internal aspect of consciousness which seems very much "hidden" from people around you. Others can see that you are conscious. For example, because you have an attentive look in your eyes or you tilt your head to hear better. However, it is seemingly impossible to know what you are conscious of, what it is you experience. The problem of explaining this is dubbed *the hard problem of consciousness* by David Chalmers (Chalmers 1996, 24). It seems to be very remote from neurons and higher brain structures. An explanation would therefore be extremely hard. In this chapter, I will dissect this idea of the hard problem further. Firstly, to discover whether Chalmers is correct in characterizing this as a problem for a science of consciousness, and secondly whether anything sensical can be said about it.

One question Chalmers asks is: Why is the experience of seeing red like *this* and not like *that*? (Chalmers 1996, 4) Why should something red not look like something blue or vice versa? Or why should seeing red not be like hearing a bird sing? An obvious and almost trivial answer would be to say that the wavelength of blue light is shorter than of red light, therefore seeing something red cannot be blue. However, such an answer would really miss the point. In answering that red light is of wavelength 480nm and blue light of wavelength 640nm, the whole notion of the experience of **red** and **blue** is gone. We are just left with a theory about factors which are supposed to underlie objective reality, not reality as we experience it. The hard problem of consciousness then does *look like* it is really deep, as it tries to get to the heart of experience itself. However, in doing so it jumps outside the range of meaningful discourse. As Hacker and Bennett point out, something red does not *look like* it is red, it *is* red (Hacker and Bennett 2022, 312). When I want to explain what red is, I will point at something red. An answer to: How does the redness of the rose look? Would be simply to point at the flower and say: Well, like this. This is simply how we can give meaning to the concept of red, by pointing to publicly accessible examples of red. There is thus nothing private about redness or other experiences, at least not in so far as we can meaningfully speak of it [1].

Then nothing is private about experience, since we can all see the same objects around us? (That is, if you are not blind.) Well, that would be too quick, I think. Surely there is some sort of subjectiveness about experience. One can wonder for example: What is it like to be a bat? As Thomas Nagel did. What would your experience be like if you could "see" with your ears like bats do? Would the world still look the same? While we might not be able to answer these questions meaningfully, by lack of being able to define any concepts related to "bat experience" properly, it does go to show that bat experience must be different from human experience, just by difference in perceptual organs. Similarly, we can ask what it would be like to be another person. For example, what would it be like to be a police officer? According to Hacker and Bennett this question is also altogether misconceived, since there are not any meaningful answers that are also interesting (Hacker en Bennett 2022, 308). The reasons are similar as to why asking why red looks the way it does is non-sensical. The only sensical responses are answers like wonderful, boring, exciting, dangerous etc. This is however not

[1] This critique is based on the private language argument of Wittgenstein. In the *Philosophical Investigations*, Wittgenstein shows that private language cannot exist, because the notions of true and false do not make sense for a language based on introspection (McGinn 1997, 129).

the thing we were getting at; it is basically like asking how one's weekend was and this surely does not give us any insight into consciousness.

However, as I said, I think Hacker and Bennett dismiss this question and the private, subjective aspect of experience too easily. There is a sense in which experience is subjective and different for everybody. For example, if you ask ten different people what they notice first in a painting like *The Night Watch* of Rembrandt, you might just get ten different answers. Another, more personal, example would be noticing the new earring my sister was wearing the other day. While I did not notice it until five hours later, my mother had seen it immediately. What this example shows is that different shapes can stand out for different people while they are looking at the *same* scene. In similar vein, I would only hear a bunch of random sounds when someone would speak Arabic to me. A native speaker would however hear proper sentences and be able to distinguish the sounds from each other and link them to meanings. More generally, it has been shown that breaks in sentences do not match up with breaks between different words in languages (Buonomano 2017). It is thus altogether impossible for an untrained listener to properly hear the words, while a native speaker will hear the words separately. Hence it is not unreasonable to suppose that people do experience the same place differently, that there is a metaphoric point of view of experience. However, whether we can meaningfully express these differences is very much contestable. Hence it is not at all certain if there can be a theory which can explain these differences in a meaningful way.

To conclude this part, the hard problem of Chalmers might seem very deep however, it is impossible to speak meaningfully about qualitative differences between people in seeing red for example. It would thus also be impossible to have a theory of consciousness about this, as it would not be possible to assert its truth, since the theory would be devoid of sense, while sense is a precondition to truth and falsity (Hacker and Bennett 2022, 467). On the other hand, it can be meaningfully said that experience is different between subjects, just not exactly how it is different.

3. Supervenience and reduction

In the coming chapter, I will take a closer look at supervenience and reduction, two closely related concepts. They are both used to explain high-*level* properties in terms of low-*level* properties. A level can be thought of as a manner of describing reality. On a low level, we can describe reality with sub-atomic particles for example. Above that we can imagine levels of atoms, molecules, cells all the way up to the level of macro reality, the level at which we can perceive the world with the naked eye.

As definition of supervenience Chalmers writes: "B-properties supervene on A-properties iff no two possible situations are identical with respect to their A-properties while differing in their B-properties" (Chalmers 1996, 33). A definition of reduction could be: "B-properties reduce to A-properties iff B-properties are A-properties" (Searle 1992, 113). The difference between these is that supervenience can leave B-properties as existing separately next to A-properties, while in reduction B-properties are ultimately A-properties. (Here and in the rest of the chapter, B-properties are 'high' level and A-properties 'low' level.) So, in reduction there is an extra metaphysical claim about the existence of B-properties (they are ultimately made up of A-properties), while supervenience is more a correlation between B- and A-properties. Reduction especially has been popular and has had great successes in the physical sciences. Think of current being nothing but electric charges moving, heat nothing but molecules jiggling.

In the same spirit, neuroscientists and philosophers have tried to use reduction to explain consciousness and conscious experience. Daniel Dennett is different from pure reductionism. He believes that the patterns of neuronal activity in the brain are the cause of consciousness. However, he also believes that it is not interesting to ask whether these activity patterns *are* consciousness. So, although Dennett does believe that consciousness is caused by purely physical phenomena, Dennett is not a reductionist, because the metaphysical question of whether certain neuronal activity *is* consciousness is just not interesting.

David Chalmers, on the other hand, has arguments to believe that logical supervenience fails for consciousness. From there he concludes that physical properties are not sufficient to explain consciousness, therefore he postulates the existence of psychical properties which underlie the existence of consciousness. This is called property-dualism. In the rest of the chapter, I will explain further distinctions in supervenience following Chalmers. At the end I will show that a reductive explanation for consciousness is different from reductive explanations in for example physics.

Logical and natural supervenience

Supervenience comes roughly in two flavours, logical and natural supervenience (Chalmers 1996, 34-35). For logical supervenience, the B-properties *always* follow from the A-properties, in every possible world. Possible worlds are alternative worlds that are logically consistent. An example of a possible world would be one where water is XYZ and not H2O. XYZ does all the same things and has the same properties as H_2O, only its intrinsic nature is different. On the other hand, it is not possible to have a world in which squares are round, since this goes against what it *means* to be a square and what it *means* to be round. Such a world would therefore be logical inconsistent and not be possible. Hence, the logical in logical supervenience gives basically a *conceptual* restriction (Chalmers 1996, 35). If B-properties logically supervene on A-properties, then we can also say that the B-facts are *entailed* by the A-facts (Chalmers 1996, 36). Here an A-fact is a state of the world which has A-properties. This state of the world can be described by a proposition.

Hopefully, an example will clarify the previous paragraph. In physics, temperature is *defined* as the mean kinetic energy of molecules. An example of a A-fact would be a description of the kinetic

energy of each molecule inside a well-defined area, let us say a box. Then a B-fact would be the temperature inside the box. And since temperature is defined in terms of mean kinetic energy, it is logically impossible that the inside of the box has no temperature. The fact that the inside of the box has a temperature is entailed by the fact that there are moving molecules inside of it. Thus temperature, the B-fact, logically supervenes on the A-fact, the description of the velocity of each molecule.

The second flavour is natural supervenience. Here the supervenience is merely contingent, B-properties in fact do supervene on A-properties, but it could have been otherwise. At least there are possible worlds in which it is otherwise. Here, we can take Boyle's law as an example. The pressure and the volume of a box with a moveable top is dependent on a constant (p*V = constant). The constant is contingent. Therefore, the exact correlation between pressure and volume is contingent too. Hence, volume and pressure naturally supervene on each other. More information (the constant) is needed to know the exact relation.

The difference between natural and logical supervenience is thus largely a conceptual issue. If the concepts, which describe the facts are not sufficiently defined, then logical supervenience is impossible. Again, for the example of temperature, now we can say that temperature supervenes logically on the microphysical facts, but 200 years ago this was not the case. At that time temperature was not defined as I described before, so it would not be contradictory for a possible world to exist in which temperature supervenes on something else than molecular movement. Thus, the supervenience between molecular momentum and temperature seems to be natural pre-definition of temperature and logical post-definition. There seems thus to be a smell of contingency in the whole difference between natural and logical supervenience. The supervenience relation changes from natural to logical, once a high-level fact is defined in terms of a low-level facts.

The other way around might also happen, in which a high-level fact was defined by certain low-level facts, but in which this relation turned out to be false. Furthermore, the high-level fact seems to stand in need of an explication, for the meaning to be well defined. Otherwise, how could we know that certain high-level facts do indeed logically supervene on low-level facts? Or that a high-level fact is reductively explained by low level facts. The low-level facts should be shown to somehow fulfil all the high-level criteria. In the next section, I will give reasons to reject the distinction between logical and natural supervenience, at least when applied to the natural sciences.

Logical supervenience problems

The first problem is the revisability problem, first raised by Quine, showing the contingency of the difference between conceptual truths and empirical truths. According to Chalmers, this is not a problem for logical supervenience, since the supervenience statement is: "If the low-level facts turn out like this, then the high-level facts are like that" (Chalmers 1996, 55). Even if empirical evidence shows that the antecedent is false, the sentence is still true (given that the high-level facts do indeed exist). So no problem, right?

However, I would argue that the sentence is incomplete. If you want to show that A-facts do supervene on B-facts, then the B-facts should at least exist. Otherwise, it cannot be shown that A-facts supervene at all. (What are they supposed to supervene on if the B-facts do not exist?) This is especially important for Chalmers, since he wants to show that all the high-level facts, except consciousness and facts related to consciousness, logically supervene on all the low-level facts + all physical laws and constants (Chalmers 1996, 71-72). However, if our description of low-level facts turns out to be false (or to change in substantial ways, which is not impossible since they are theoretical), then he has not shown anything! He will not have shown that macro facts logically

supervene on something, since the facts on which they were supposed to supervene do not exist[2]. So, I argue that Chalmers needs the truth of the low-level facts to be able to use the supervenience relation. Without it, the supervenience relation becomes useless. Therefore, the revisability problem does pose a problem for the difference between conceptual and empirical truth, or between logical and natural supervenience.

Furthermore, logic can be used inside a theory. With valid deductions, we can make predictions (propositions) about reality. These propositions can be true or false. For example, from the theory of relativity we can logically deduct that the path of light is curved around large objects. The truth or falsity of this deduced proposition is determined by empirical tests. So, the truth of a prediction does not necessarily follow from a theory. However, with logical supervenience, B-facts necessarily follow from A-facts. Then given our incomplete physical theories, I do not believe that is appropriate to use arguments form logical supervenience to support statements about reality. Because how can we assert the implied *necessity* in logical supervenience when our knowledge about (physical) reality is incomplete?

For a priori entailment of B-facts by A-facts, explicit analysis of the B-facts in terms of the A-facts is not necessary according to Chalmers (Chalmers 2010, 213). Generally, it is enough to have an *intension* about the meaning of the B-facts, a "function specifying how a concept applies to different situations" (Chalmers 1996, 54). However, this intension is derived from intuition, which can be thought of to be based on experience, which ruins the a priori part. Furthermore, intuition is not necessarily correct. Think of results of modern physics (no well-defined present, no well-defined momentum position etc.) as examples where old intuitions were shown to be false. This all shows that it is doubtful that a priori entailment is possible, and even if it is, that it is not so useful in creating any certainty.

Logical Supervenience of the Macroworld

Chalmers believes that the macro world (except for consciousness and things dependent on consciousness, as we will see in the next chapter) logically supervenes on the micro-physical facts (Chalmers 1996, 37). According to Chalmers, this means also that almost everything can be reductively explained by microphysical facts and physical laws (Chalmers 1996, 48). Logical supervenience supports the metaphysical claim of reduction, because given the description of the A-facts, the description of the B-facts will always follow logically. In such a case there is a strong suggestion that the B-facts is indeed nothing but the A-facts.

Chalmers conceives that it is impossible that given all the microphysical facts, the macro world will be different in any possible world. Since it is inconceivable, it is logically impossible. Therefore, the macro facts supervene logically on the microphysical facts (Chalmers 1996, 71). However, I do not think that this is the case. I can for example imagine a world with the exact same microphysical facts, but in which time runs backwardly. All the microphysical facts will still be at the same time and place, but their *order* will be different. This will be a radically different world. Furthermore, I can conceive that the microphysical facts are the same, but that the causal interaction between them will be different, giving rise to different high-level facts.

To prevent problem such as these, Chalmers also adds all the physical laws (Chalmers 1996, 75). Our current physical laws are imperfect, both in completeness and comprehension of them. However, this is not necessarily a problem. All that is necessary is that these laws exist, not that we know them.

[2] In his analysis, Chalmers takes the microphysical facts for granted, to bypass sceptical doubts (Chalmers 1996, 75-76).

When the same physical laws are in place with the same microphysical facts, then it seems indeed to be the case that all the high-level physical facts are entailed.

Chalmers thus makes a metaphysical commitment to the existence of physical laws and the existence of microphysical facts. But he cannot use these physical laws to establish that consciousness does *not* logically supervene on the physical. Whether or not consciousness is fully dependent on the physical needs to be established by a separate argument, because it does not just follow from the existence of physical laws and microphysical facts. It would only be possible if we would *know* the complete set of physical laws and microphysical facts. In that case, we could in principle be able to see whether or not consciousness follows from that set. However, this is *not* the case, so Chalmers will need a separate argument to show that consciousness is not logically supervenient on the physical.

Reduction and consciousness

In the final part of this chapter, I will explain why consciousness cannot be reduced in the same way as other concepts and why this is not strange; it should be expected from the very idea of reductive explanation. In explaining macro phenomena, reductive explanation tries to find the underlying causal base of the phenomena, to find the objective reality underneath the subjective appearance. Subjective appearance of heat might be different for various people, I might find something extremely hot while someone else would not. The reductive account of temperature and energy transfer is able objectify this by using molecular momentum to define temperature and temperature gradients between bodies to calculate the energy transfer between them. In the reductive explanation of heat there is thus no need for the subjective appearance of heat, so reduction enables the separation between objective reality and subjective appearance (Searle 1992, 118-124). The objective reality is explained, while the part of the subjective reality is kicked further down the road.

This immediately shows that a reductive explanation of conscious experience is different, because in this case the appearance *is* the reality (Searle 1992, 118-124). A reductive explanation of conscious experience cannot exist without the subjective appearance. Therefore, such an explanation needs to make a metaphysical claim, that a certain objective state causes or is a certain conscious state. Hence, a reduction of consciousness is different in so far as a "normal" reduction goes from objective facts (e.g. molecular momenta) to objective facts (temperature), while consciousness goes from subjective facts (e.g. a conscious experience) to objective facts (e.g. activity of certain brain areas). The metaphysical load of the latter is arguably higher.

Dennett argues for an explanation of consciousness based on unconscious, physical events (Dennet 1991, 454-455). An explanation of pain consists then in certain activity happening in the brain and the body. However, Dennett is not claiming that this activity *is* pain, thus he does not want to make that metaphysical claim. But if the activity in the brain is not the pain, has pain then be explained? I would argue that it has not been explained by Dennett, he merely describes a supervenience relation between pain and physical events, which is not sufficient for an explanation. It does not explain *why* certain brain activity causes pain experience.

In this chapter, I have explained reduction and supervenience, and the ways in which David Chalmers uses them. I have set out arguments against the use of logical supervenience to support any statements about reality. Furthermore, I have discussed the logical supervenience of high-level facts on the microphysical facts. Finally, I have tried to show that a reduction of consciousness will be different than a reduction of a macro physical process, the former reduction will have a higher metaphysical cost.

4. Zombies and non-physical properties

Chalmers' view

In this chapter, I will discuss the zombie thought experiment and possible conclusions. The idea of the zombie thought experiment is to imagine a world which is an exact microphysical duplicate of ours and which has all the same physical laws. So, in this world, there will also be people walking around, doing their daily business etc. Now the big question is whether these people are also necessarily conscious. According to Chalmers, it is at least conceivable that these people are not conscious in so far as they do not have conscious experiences (they are zombies). They might perhaps act as if they are conscious, they can still talk about what they see for example without actually having the visual experience. Hence, consciousness, in so far as it involves experience, does not necessarily follow from the microphysical facts. This shows according to Chalmers that consciousness is not logically supervenient on the microphysical, it is only naturally supervenient. Therefore, it requires further facts to be explained, some sort of (proto-)consciousness properties of particles (Chalmers 1996, 93-99). This would ultimately allow consciousness to be logically supervenient on these new properties and thus to be reduced upon those (Chalmers 1996, 126-129). Of course, Chalmers has also other reasons for stipulating these new properties, but for reasons of length I cannot discuss these here.

Before dissecting the zombie argument further, I first want to jump ahead to the conclusion Chalmers made, which feels a bit peculiar. Firstly, it makes Chalmers accept the Chinese nation thought experiment (see below for details) which is at first glance strange. Furthermore, the only way in which we have experience (as far as we can tell) of these (proto-)consciousness properties is by experience itself. However, if they are basic properties like as charge is, they might have widespread effects, which we cannot notice or of which we do not know that we notice them. Either of these options should be explained by a developed theory. Of course, these peculiarities can also be seen as interesting discoveries. Therefore, they cannot be counted as a proper counter argument against Chalmers' conclusion. However, there are reasons to dismiss the thought experiment in the first place.

Circularity?

Firstly, the stipulation of new properties rests largely on the fact that Chalmers thinks he can assign consciousness a special place as a non-reductive property in a world in which everything else is logically supervenient on the microphysical facts and physical laws. When Chalmers conceives of the zombie world, he conceives of a world with the same microphysical facts and the same physical laws as in our world. But, in this world there will not be any consciousness, zombies do not have phenomenal experiences. From this he concludes that there are certain psycho-physical laws and properties which are active in our world, causing consciousness. However, they do not exist in the zombie world, causing the absence of consciousness.

The requirement for the validity of the thought experiment is that the conceived zombie world is logically consistent. For example, it does not matter whether we *know* the physical laws, this is an epistemic question. What matters is that it is logically consistent that consciousness is *not* entailed by physical laws and microphysical facts. The thought experiment runs like this:

1. There is a possible world which is physically identical to ours.
2. It is conceivable that there is no consciousness in this world.

Consciousness is not logically supervenient on the physical.

But then we need to ask the question: Why is it conceivable that there is no consciousness in the physically identical world? This needs to be established logically, since a conceivable situation needs to be coherent. Furthermore, the reasons need to be independent of physical laws and the microphysical facts, as I wrote before. However, Chalmers does not provide a clear answer, he writes: "A zombie is just something physically identical to me, but which has no conscious experience - all is dark inside. While this is probably empirically impossible, it certainly *seems* [my italics] that a coherent situation is described; I can discern no contradiction in the description" (Chalmers 1996, 96).

Chalmers thus has the *intuition* that consciousness is not entailed by the complete set of microphysical facts and physical laws. This is problematic, because it runs the risk of circularity. The intuition is perhaps based on a prior assumption that it is not possible to explain consciousness from the physical facts. Without this assumption, it is hard to see why it is logically possible that consciousness is not entailed by the physical, while all the other high-level facts modulo consciousness *are* entailed by the physical. What can be the logical basis of this distinction between consciousness and all the other high-level facts? However, if this assumption is indeed being used here, then the whole zombie thought experiment becomes circular.

Meaning of consciousness

If the assumption is not used, on the other hand, then the case for the conceivability of zombies is still not clear-cut. Intuitions are not necessarily correct as I pointed out earlier, they can misrepresent possibilities. Furthermore, perhaps the concept of consciousness is insufficiently clear, resulting in an apparent possibility to conceive of zombies, without it being possible to do so. If this is the case, then the argument does not show anything, because the concept of consciousness just lacks sufficient meaning to follow logically. I already alluded to this problem before in my discussion on logical supervenience.

Additionally to the insufficient clarity of the concept of consciousness, there is the problem of whether phenomenal concepts can have proper meaning. In the chapter on the hard problem of consciousness I showed that the meaning of words describing experience, such as **red**, cannot be established privately. Hence, the meaning of a *phenomenal quality*, in the sense that the quality is purely private, cannot be established. However, in the thought experiment, it is phenomenal qualities that the zombies lack. If the meaning of phenomenal qualities cannot be established, how can they then be used in a thought experiment? If a concept is not definable, then how can the concept be used in any sort of meaningful way?

Conceivability

Next, Chalmers bases his argument on the conceivability of zombies in a physical identical world. According to him, conceivability entails logical possibility. However, he cannot actually prove this (Chalmers 1996, 68). His only defence of this thesis is by countering counterexamples. But this cannot be enough. Just because there is no counter example against the generalized continuum hypothesis, does not mean that it can be taken for granted (Hill and McLaughlin 1999). Therefore, mere conceivability of zombies does not need to mean that they are logically possible.

Another counter argument against Chalmers is that the Goldbach conjecture (an unproven, mathematical idea) can be conceived to be true or conceived to be false. If metaphysical possibility follows from conceivability, then both the truth and the falsity are possible situations (Chalmers 2010, 145). But logically, only one of them is an actual metaphysical possibility. According to the conceivability argument, both of them are metaphysically possibilities, so we end up in a

contradiction. Logically and thus metaphysically, only one is possible, but according to conceivability, both are possible. Thus we must conclude that conceivability does not entail metaphysical possibility.

Chalmers counters this argument by stating that an ideal conceiver would be able to actual tell if the Goldbach conjecture is true or false. Therefore, the above is no problem according to Chalmers. Ideal conceivability can still entail possibility. Ideal conceivability "abstracts away from those [the subject's contingent cognitive] limitations" (Chalmers 2010, 143). If it is possible to ideally conceive of logically coherent situations, then these situations might indeed be possible in an alternative world (Chalmers 2010, 146-147). However, how is ideal conceivability possible? How can one abstract away from one's own cognitive limitations? If Chalmers wants to use the conceivability of zombies, then he should show how or why he is able to ideally conceive them. Otherwise, he has not shown anything.

A sceptical problem

Finally, there is a problem with the logical possibility of zombies. What Chalmers conceives of is that zombies do not have experience as we do, they just fake it. When they see red, they do not actually have any conscious experience of red, but they do say that they see red. This is possible because the functioning of the eyes and necessary mechanisms in the brain supposedly are explainable by the microphysical facts. The question then is: How can we know then that these beings are not among us? We cannot tell the difference between someone seeing red and someone acting as if he sees red. If zombies can metaphysically exist, then we cannot exclude them from our world either, which is rather absurd (Hacker and Bennett 2022, 362). If consciousness is dependent on some undiscovered properties, then we cannot tell which human is actually conscious and which one is like a zombie, because we do not know how these properties give rise to consciousness. Acceptance of the zombie thought experiment thus leads to an absurd situation.

Dennett's view

In the final part of this chapter, I will go further into Dennett's reaction on the zombies. In short, he thinks that the zombies are conscious. They can express in words what they think they experience. So, they have internal states, on which they can report. Hence, they can be said to be conscious of their internal states (Dennet 1991, 311). It would thus at least seem to the zombie as if it were conscious. This seems to be some sort of illusion, a User illusion as Dennett calls it (Dennet 1991, 311). However, we humans suffer from the same illusion according to Dennett. Therefore, zombies are as much conscious as we are.

The User illusion refers to the gap between what you see on your computer screen and what is actually creating the images on the screen. When you do not know anything about computers, this entire process may seem like some sort of magic. However, as most people know, a computer is just a bunch of electrical circuits which are manipulated in an orderly fashion, ultimately creating practical user interfaces. The same supposedly happens in the brain, while all the neuronal circuits may seem totally disconnected from your conscious experience, they ultimately cause it. It is an illusion however that all the input from the sense comes together at one point from where we can judge about experience. Just like it is an illusion to think that the computer screen is necessary for the computer to operate (Dennet 1991, 312-313).

The complete phenomenological experience becomes then an illusion, some sort of rational construct. But then why do we have this illusion? Just to be able to report upon received sensorial input? Also, what causes the illusion? Dennett has fought hard to discharge the idea of a Cartesian Theatre, a place in the brain where all the sensorial input comes together (Dennet 1991, 101-138). So, the illusion is certainly not caused by something like that, but then what does cause it? Claiming that it just *seems as if* I have this complete, continuous experience (Dennet 1991, 366) is not

sufficient, then you also need to know why and how this seeming experience exists. However, this is what Dennett seems to do. Kicking the hard problem of consciousness down the road by explaining it as an illusion. But an illusion everybody seems to experience nonetheless.

Overall, I would say that Dennett has not fully explained consciousness. Dennett tries to explain away conscious experience by trying to explain it as an illusion. However, this is not convincing. As a result, Dennett's view that zombies are conscious cannot be upheld within his framework, since they can only report about their internal states, without having experience. This is arguably insufficient for consciousness.

In this chapter I presented the zombie thought experiment and what conclusions Chalmers and Dennett take from it. I presented arguments against the viability of the thought experiment itself, the arguments Chalmers had for his conclusions and the implications of his conclusions. So here I rest my case against Chalmers and his dualism. I ended by shortly discussing Dennett and his view on the thought experiment. His theory is at least incomplete and that his argument for the consciousness of zombies is insufficient. In the next chapter, I will investigate computationalism.

5. Computationalism

In the previous chapters, the focus has been on expanding Chalmers' reasons for his dualism and showing why he is mistaken in believing that extra (non-physical) properties are logically needed to explain consciousness. Here, I want to focus on a reductionistic explanation: computationalism. First, I will shortly present how computationalism tries to explain consciousness and then I will show why it is not a complete explanation with some of Searle's arguments.

The main idea behind computationalism is that the brain consists of computational units which process sensations and perceptions in order to give (behavioural) outputs (Searle 1992). So mental states, and thus consciousness, are reduced to computations. The mental states can then be described in some form of information processing by small units which perform some sort of computation. A simple example would be a heat reflex. Imagine you touch a hot stove by accident. The moment you to touch it, you pull your hand away. This is explained by saying that the heat receptors in your fingers are activated. Subsequently, the sensory neurons signal onto the motor neurons and finally the motor neurons activate relevant muscles which creates the withdrawal from the stove. In terms of computationalism, the input is the activation of the heat receptors, the "computation" is done by the sensory and motor neurons. Now the idea is that consciousness and its functions can ultimately also be described in terms like these. Even more, if this idea is true, the strong AI is very much possible and we can expect to have conscious machines in the future.

The causality of information states

In computational explanations, neurons can be described as processing information and changing information states. Let us delve a bit deeper into this idea. The mathematical idea of information was originally developed by Shannon around seventy years ago. The theory was purely syntactical, so no semantics at all, and it put the specificity of states into a mathematical framework, given the range of possibilities for that state (Chalmers 1996, 278). A state in this case can be anything, for example four bits. Each bit has two possibilities and there are further combinatorial possibilities. Information states can also be created in the real world, for example on any CD. The CD surface consists of pits and lands, which reflect the light back differently, creating two possible states. So the CD can be described as an ensemble of 0's and 1's.

Finally, instead of being discrete, information states can also be continuous. So in the first example, instead of assigning 0's and 1's, we could have assigned any number between 0 and 1. These information states are purely syntactical, so they do not have any intrinsic meaning. Meaning is ascribed by some agent. The CD can only be said to cause music to be played because we ascribed meaning to the pattern of pits and lands on it. This is important to keep in mind, information is a construction and so is the possibility of reading out information.

Based on our experience we can also *construct* information spaces. For example, colours can be described in a 3-dimenional continuous information space (type colour sphere in your search engine if you want to see it). Sounds would be a bit more complex, but they should also be describable as some high-dimensional continuous information space (tones and intensities of tones). Now why would all this information talk be interesting? The basic idea seems to be that phenomenological information spaces are somehow realized by physical systems in the brain (Chalmers 1996, 284-287). So all we need to do then, to understand which parts of the brain are causally relevant to our phenomenological experience, is to find ensembles which have the same information space, which "code the information". So far so good it seems. However, there are some problems around this, so we should tread carefully.

Firstly, the brain cannot be said to "code" any information (Hacker and Bennett 2022, 328), as code is a purely syntactical business. We, as semantic agents, give meaning to code (Searle 1992, 225). For

example, when I write a program on my computer, I give meaning to it by deciding what input it can take and what output it needs to generate. I could explain the outcome of the program by saying that it is caused by the code. This would be a teleological explanation, since it appeals to the design of the code. Teleological explanations can be useful in cases where we know that there is a designing agent. However, for consciousness we would need to postulate this agent, which would really beg the question. So unless we have reason to believe that some sort of God designed our "brain program", we should stay away from these sorts of computational explanations.

Now this might seem like dead-doing argument, one could say that this is all obvious but that it provides a nice framework in which we can explain the workings of the brain. However, this too is misconceived. The information state ascribed to a brain process itself is not causally relevant in any brain process, it is merely a description (Searle 1992, 224). What is relevant is the physics. A neuron does not start signalling because it "wants" to create a certain information state. No, it starts signalling because of some physical properties. Another way to see this difference is with weather simulations. Do the weather simulations we run to predict the weather *explain* why the sun is shining or why it rains? I do not think so. The shining of the sun is explained by absence of clouds, which has a host of underlying physical reasons. Rain is explained by water vaporizing on the ground, going up in the air, condensing into droplets and eventually falling again. These are all physical reasons. Hence, a computational explanation would really not be an explanation at all, it would only be a description.

A Chinese thought experiment

A thought experiment may help to show how strange and incomplete an explanation in terms of information processing would be. It is called the Chinese Nation, conceived by Ned Block (Block 1978). Imagine that a group of people (the Chinese population) would implement the exact same signalling process as neurons do in the brain. Exactly how is unimportant, what matters is that from an information point of view both systems, the human and the neuronal one, act the same. If being in pain or seeing an object are just certain information states, then their implementation would not matter. So if the information states of being in pain or seeing red would be implemented on the human system, then the Chinese Nation should be also in pain or have an experience of red, somehow. However, this seems to be rather odd. Furthermore, it is altogether meaningless to ascribe pain to a system comprised of so many people. Pain is something that applies to individuals, not to a set of individuals. The point of the whole thought experiment is that conscious experiences cannot be explained by an information processing theory or otherwise we get strange statements as: the set of one billion people is in pain, without any individual being in pain. Adhering to a theory in which such things can be said seems to be accepting a non-sensical theory. Hence also computationalism should be rejected.

So information itself should have no role in a final explanation of any brain processes, thus also not for consciousness. However, what it can do, is to point out areas of interest in the brain. To go back to the colour example, it can reasonably be assumed that an ensemble of neurons causally involved in vision should be describable by a similar 3-dimensional information state. Any structure that does not have this property can be set aside already. So the role I would envisage for information and computationalism is merely an exploratory role, to be later swept aside by physical explanations; similar to how AlphaFold can predict protein shape and function by simulation (Jumper, et al. 2021). The prediction is however not real, actual empirical experiments are needed to find out the actual shape and functioning of the protein. AlphaFold is thus useful to narrow the scope of possibilities of proteins of interest at the start of research. At the end you would like to have a physical explanation, the same applies to computationalism.

6. Consciousness beyond experience

In the previous chapters, I have looked at different ways of explaining conscious experience. I have argued that Chalmers' reasons for postulating some new sort of consciousness property are mistaken. Therefore, his property-dualism can be rejected, until there actually are good reasons to accept it. Furthermore, I have shown that a computational explanation is also not fully apt for explaining conscious experience. Finally, I suggested that a reductive explanation of conscious experience comes with the metaphysical load of having to bluntly accept a certain correlation between physical and mental states. However, one must wonder whether this metaphysical assertion is meaningful. In this chapter, I will suggest that neuroscience should study consciousness by finding the material conditions in the human body for the human to be conscious.

If we were to take these conditions as an explanation for consciousness, then we would not actually have explained consciousness, we would merely have redefined consciousness in terms of its physical correlates. Things like conscious experience would really not follow from such a redefinition, it would be lost. Maybe it is also not meaningful to expect this from an explanation. When I ask the question: Why do you see a red apple? You can just answer by saying: Because it is in front of me (Hacker and Bennett 2022, 343). An answer involving the retina and subsequent neuronal action will not be necessary. Such an answer would be useful when the question is: What are the underlying material conditions in the brain which allow one to report on what he sees?

Furthermore, I would also not call these material conditions causes of consciousness or conscious experience. It is really the totality of conditions around us that cause us to be conscious. You are conscious of something, why? Because it made a loud noise or because it has bright colours etc. These sorts of explanations are fine. Describing all the physical conditions in your brain really would not show why you were conscious of that thing. The physical conditions might show why certain neurons started sending signals and others not, but it would not show that you perceived some colour or sound. Conscious experience then seems to be something that is beyond our science to explain. It shows itself; I can see that someone is conscious.

Besides looking for neurological conditions, we can study consciousness by how we use it in our language, we can study its meaning. For example, the meaning can be showed by grammatical investigation and conceptual analyses such as applied by Hacker and Bennett in the Philosophical foundations of Neuroscience (Hacker and Bennett 2022, 338-342). In their analyses, they show how parts of consciousness are really more like abilities, which rely on certain conditions to be exercised. However, it really is not meaningful to state that there are abilities in the brain. What is in the brain are just the structures which enable the exercise of certain abilities when the outside world causes their use. Seeing red is then caused by seeing a red object and enabled by certain material conditions in your brain. In this manner, there does not need to be a brute metaphysical claim which asserts the causation of certain mental states by certain brain states.

7. Conclusion

In this thesis I have mainly looked at two approaches which ultimately aim to explain consciousness: property-dualism as Chalmers uses it, and computationalism and reduction. I rejected property-dualism because there is not sufficient reason for postulating the existence of some unknown properties and the existence of them might lead to some rather absurd situations. Computationalism on the other hand loses sight of consciousness in its explanation mode. In redefining consciousness or conscious experience into some neuronal(network) properties, the consciousness is not explained, it is merely corelated with neuronal activity. Therefore, I propose that neuroscience searches for the physical conditions for humans and other animals to be conscious, without claiming that these conditions cause or are consciousness. They form the necessary physical basis. So ultimately, I believe that some physical model cannot explain consciousness.

We can perhaps get a greater clarity on consciousness by conceptual analysis as Bennett and Hacker propose. For example in how we use consciousness as a word in language. Its meaning is shown in the numerous ways in which we use it. Also we can see that people are conscious, it really is not some mysterious force, rather it becomes mysterious because we try to say what consciousness is. However, it rather seems to be at the boundary of what we can meaningfully say in language. Similar for conscious experience, we cannot know or say that other people experience the world differently. However, we can show these differences in paintings, literature, poetry and music.

Works Cited

Allen, Colin, and Marc Bekoff. 2017. “Animal consciousness.” In *The Blackwell Companion to Consciousness*, by Susan Schneider Max Velmans, 63-76. New York: John Wiley & Sons Ltd.

Bachelor, Alexandra, and Jean-Robert Roussel. 2001. “Altered State and Phenomenology of Consciousness in Schizophrenia.” *Imagination, Cognition and Personality* 141-159.

Block, Ned. 1978. “Troubles with Functionalism.” *Philosophy of Science* 261-325.

Buonomano, Dean. 2017. *Your Brain Is a Time Machine: The Neuroscience and Physics of Time.* New York: WW Norton & Co.

Chalmers, David J. 2010. *The Character of Consciousness.* New York: Oxford University Press.

—. 1996. *The Conscious Mind.* New York: Oxford University Press.

Dennet, Daniel C. 1991. *Consciousness Explained.* London: Penguin Books.

Gulick, Robert Van. 2021. *Consciousness.* 21 December. Accessed May 13, 2022. https://plato.stanford.edu/archives/win2021/entries/consciousness/.

Hacker, Peter M. S., and Maxwell R. Bennett. 2022. *Philosophical foundations of Neuroscience, 2nd edition.* Hoboken: John Wiley and Sons.

Hill, Christopher S., and Brian P. McLaughlin. 1999. “There Are Fewer Things in Reality Than Are Dreamt of in Chalmers' s Philosophy.” *Philosophy and Phenomenological Research* 445-454.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly accurate protein structure prediction with AlphaFold.” *Nature* 583-589.

Kandel, Eric R., James H. Schwartz, and Thomas M. Jessel. 2012. *Principles of Neural Science.* New York: McGraw-Hill.

McFarlane, William R. 2011. “Prevention of the First Episode of Psychosis.” *Psychiatric Clinics of North America* 95-107.

McGinn, Marie. 1997. *Routledge Philosphy Guidebook to Wittgenstein and the Philosphical Investigations.* London: Routledge.

Searle, John. 1992. *The Rediscovery of The Mind.* Cambridge: The MIT Press.