Does Free Will Matter?

Agent Causation and Qualia Fusion Emergence

Camilla Karlsen

Thesis presented for the degree of

MASTER OF PHILOSOPHY

Supervised by Professor Carsten Hansen

Department of Philosophy, Classics, History of Art and Ideas

UNIVERSITY OF OSLO

Høst 2020

Does Free Will Matter? Agent Causation and Qualia Fusion Emergence.
Camilla Karlsen
http://www.duo.uio.no

Preface

During my philosophical journey I have become more and more fascinated by conscious experience. We all have this wonderful access to reality through our perception and it seems obvious that we, as conscious agents, interact freely with the material world. This thesis is an attempt to show that agent causal accounts of free will are not as mysterious and incomprehensible as it now appears to science. Fusion emergence, and the examples from quantum physics that contribute fascinating and surprising new insights into the nature of matter, might help provide the first steps towards a scientific account of free will.

I want to thank my supervisor Carsten Hansen for all the helpful comments and suggestions throughout the process of writing this thesis. I also want to thank my family for their patience, and my dearest Christian, for the endless support and wonderful conversations.

Abstract

In this thesis I consider how agent causation can be reconciled with a broadly physicalistic framework. I suggest that agent causation is a fusion emergent causal power. The theory of fusion emergence due to Humphreys, recently exemplified by phenomena discovered by Aharonov and colleagues in quantum physics, opens up the possibility that phenomenal consciousness is fusion emergent. The irreducible nature of qualia (phenomenal properties) suggests that they might be understood as fusion emergent phenomena. If we consider qualia as fusion emergent phenomena with downward causal power, it seems that we can make room for a genuine physical explanation of free will, able to avoid epiphenomenalism, overdetermination and dualism between the mental and physical realm.

Contents

Introduction

"This is the highest wisdom that I own; freedom and life are earned by those alone who conquer them each day anew."- Johann Wolfgang von Goethe

In our daily lives, most of us think we have free will. It simply seems obvious that my choice between coffee or tea this morning was a free choice, not predetermined by natural laws or anything else. However many scientists and philosophers today would say that this experience of freedom is just a comforting illusion, that in fact everything we do is predetermined by the course of nature.

The agent-causal account of free will has not been popular in philosophy in recent years because it seems to go against the scientific and physicalistic worldview that most thinkers hold as true. For many philosophers it seems simply incredible that humans can have the causal power needed to act freely in the way we think we do. My aim in this thesis is to give a metaphysical conception of how agent-causal libertarianism can make sense within a broadly physicalistic framework. If this picture holds, agent causal libertarianism might not be as weird as it may appear to be with all the philosophical assumptions we make.

In order to situate my position in a broader context I will begin by contemplating the different attempts to make sense of free will by various philosophers throughout time. I give a brief overview of the three main positions in the free will debate in the first chapter, namely strong determinism, compatibilism and libertarianism. The second chapter is a short introduction to libertarianism, where I chose to defend agent-causal libertarianism, because it is the only position that can give us the robust sense of freedom that enables us to choose our actions and control our physical bodies and surroundings independently of any predetermined happenings. In the third chapter I go deeper into the problem of mental causation in a material world and argue that only strongly emergent mental properties can have the causal power needed for free action. This is because weak emergence does not seem able to give a plausible account of how mental causation can happen independently of its micro-physical realizers.

In the last three chapters I develop my own view, where I suggest that there might be a connection between the new philosophical notion of "fusion emergence" developed by Humphreys (2016) and agent causation. It looks like fusion emergence can give us the radical kind of downward causation needed to explain mental causation. The elements involved in the fusion process seem to disappear and thereby give rise to an entirely new property that cannot be reduced to its parts. This may well sound surprising and strange, but we also find support from the eminent physicist Aharonov and colleagues (2018), who demonstrate a clear example of a kind of fusion emergence in standard quantum physics. In chapter four I will explain fusion emergence in more detail. In chapter five I argue that mental causation must be conscious in order to count as free, and I draw the connection between fusion emergence and qualia, as both are irreducible and seem to imply downward causation (that I believe must be compatible with a broadly physicalist theory). In the last chapter I suggest that the conscious agent is essentially the epistemic self, consisting of a unified qualia-space, able to act freely with downward causal power. We see that fusion emergence, phenomenal consciousness and the acting agent combined enables us to conceive of the possibility that free will can matter.

Why is free will important?

In this thesis I argue that free will is possible. But does it really matter whether or not we have free will? Why is free will important? There are many reasons for finding free will important. If we are free to choose our goals and actions we may also be more responsible and thoughtful, as well as creative in trying to find solutions to problems. It can make life more meaningful, in the sense that if one has choices that matter, this means that one can have a real impact on the world. In another sense it is also related to the question of moral responsibility; if we are not responsible for our actions, it is more difficult to argue that we can be held accountable for the harmful things we do.

The importance even of just *belief* in free will may also be important. For instance, empirical studies have shown that believing you have free will actually makes you act more morally and responsibly. Controlled studies have shown that the subjects of the experiments who are given arguments against free will, like the argument that free will is an illusion or that the mind is epiphenomenal, are more likely to lie and cheat than subjects who are given arguments that defend free will (Stapp 2008, p. 4). Studies have also shown that beliefs about

free will can change brain processes related to a very basic motor level (Rigoni et al. 2011, p. 613). This shows that our beliefs about free will have important consequences for our lives. In particular, it seems that if people disbelieve in free will, this belief alone can have actual negative consequences for their lives.

This is not to say that we should aim for anything less than the truth, and nothing but the truth, about free will, regardless of the consequences. If it turns out that there is no free will, then that's just how it is, and we just have to face the negative consequences of this view growing in popularity, and find ways to deal with it. However, the possibility is still quite open. And so long as the question is open we should be careful to proclaim that free will is impossible unless we are relatively certain about it. We are evolving our knowledge of ourselves and the universe rapidly, and declaring defeat for free will seems premature.

Perhaps the greatest challenge for the possibility of free will is understanding how the mental can possibly have any control over physical objects. For example, how can the following thought; "I will pick up this cup", cause my physical hand to pick up a physical cup? From a pretheoretical view, it seems like a rather plain and obvious truth. However, from a scientific standpoint it seems reasonable to think that the physical world is "causally closed" which means that every physical effect must have a sufficient physical cause. If this is true, then it is hard to make sense of how our thoughts can cause something physical to happen, unless the mind itself can be reduced to its physical parts (Robb and Heil 2019). The weight of science as an authority in the last several hundred years, makes some believe that this is the nail in the coffin for free will. A more sober view may be that while our theories seem to present a challenge against our pretheoretical sensibilities, it is important to not overestimate our understanding of nature, or to underestimate the ability of nature to repeatedly confound us.

Chapter 1: Overview of the free will debate

There is apparently a large gap between our intuitive experience of free choice and a completely deterministic universe. We have a strong intuition that we sometimes could have thought or acted differently than what we just did, which gives us an experience of having a choice in the matter. But what does a "free choice" consist in? Different thinkers have given many different answers to this question, so I will begin by presenting an overview of the free will debate in this chapter.

The Stanford Encyclopedia of Philosophy defines free will as a significant kind of *control* over one's actions and that this control consists mainly in the "freedom to do otherwise" and "self-determination" (O'Connor and Franklin 2019). Let's take this as our initial definition of free will as we survey the different views people have on this topic.

Incompatibilism and compatibilism

One of the main questions in the free will debate is whether or not free will is possible within a determined universe. Incompatibilists answer "no", while compatibilists answer "yes" to this question. Incompatibilism is the view that free will is incompatible with a completely determined universe. If everything really is predetermined, it is hard to see how our choices can be independent of the predetermined course of nature. There are many different theories that try to grapple with this problem.

The main positions I will focus on in the free will debate can be summed up in these three positions; Hard Determinism (free will is incompatible with determinism and therefore free will is impossible), Libertarianism (free will is incompatible with determinism, but determinism is false, so free will is possible), and Compatibilism (free will is compatible with determinism and therefore is possible within determinism).

After a short historical background, I will begin by presenting an overview of hard determinism and then compatibilism. Libertarianism is the main topic for Chapter 2, and will be treated more thoroughly, as this is the view I argue for in this thesis.

Historical background

Throughout history, philosophers have had much to say about free will. Most of their theories on this subject can be fitted into one of the main positions; hard determinism, compatibilism or libertarianism.

For Plato free will is a kind of self-mastery that can be achieved by developing the virtues of courage, wisdom and temperance which leads to a liberation from base desires and impulses and gives us a better understanding of the "Good" (O'Connor and Franklin, 2019). This can be understood as a form of Libertarianism, because self-mastery seems to include the power to choose our actions.

Aristotle shares much of Plato's views and says that humans as rational agents have the power to choose, and much of our actions are voluntary. We can be the cause of our actions and we can be aware of the circumstances of our actions. Mature humans can make choices after deliberating different means to our ends from rational principles of action. If we consistently choose well, we will develop a virtuous character that will form over time and it is in our power to be either virtuous or vicious (O'Connor and Franklin, 2019).

The stoics believed that all human behaviour and choice was causally determined but that this was compatible with our actions being "up to us" if they come about "through us". (O'Connor and Franklin, 2019). This might be labeled as an early form of compatibilism.

Augustine thought that the will is a self-determining power and no powers external to it can determine its choice, but he did not rule out that the will can be internally determined by psychological factors. On the other hand Augustine had theological reasons to think that all things are determined by God, so scholars disagree on whether Augustine was a libertarian or a compatibilist. He did think that we are affected by desires that make it impossible to will something that goes against those desires. This can keep us from gaining "true freedom", which is only possible when our will is aligned with the Good (O'Connor and Franklin, 2019).

Spinoza can be said to have been a hard determinist. He held the view that everything is categorically necessary, opposed to the weaker form of necessity endorsed by most

determinists. He argues that there is no room for free will either for humans or for God. But he did not think that the absence of freedom would have any terrible consequences. For Spinoza, a kind of self-determination happens when our feelings are determined by true ideas of what is real - when we desire nothing but truth and when the better part of us is in harmony with the whole of nature. Spinoza was the forerunner of many later free will sceptics and this is a position that still has strong support (O'Connor and Franklin, 2019).

We see that some philosophers fit pretty easily into one of the three main positions, while there are disagreements about how to place other historical figures, like in the case of Augustine. Overall there is no doubt that most great thinkers throughout history have taken the question of free will seriously and that it is still one of the most discussed questions in philosophy.

Determinism

Stanford encyclopedia of philosophy states a general definition of determinism as follows "*The world is governed by (or is under the sway of) determinism if and only if, given a specified way things are at a time t, the way things go thereafter is fixed as a matter of natural law*" (Hoefer, 2016). This kind of definition is in line with the general framework of contemporary science, using scientific laws to derive consequences from initial conditions of a physical system. That's also partly why it has such force, because it is aligned with the success of modern science in observing, understanding and manipulating the world around us with great precision (Hoefer, 2016).

Determinism has often been confused with fatalism. Fatalism can be divided into two main categories, theological and logical. Theological fatalism is the thesis that events are destined to happen because of the will of God or some divine foreknowledge. Logical fatalism is the idea that since something is either true or false, future events are already determined (since the fact that they will happen is either true or not). Determinism differs from fatalism in that events are fixed, not by God or logic but by natural laws or cause-effect relations. Determinism and fatalism do agree in the assumption that given the way things have

happened in the past, all future events that will happen are already destined to occur (Hoefer, 2016).

Arguments against free will

Philosophers who deny the existence of free will completely, are often called Hard Determinists. The most radical a priori argument is that free will is not merely absent, but completely impossible. Galen Strawson argues for this and he associates free will with being "morally responsible" for one's actions. According to him we cannot be responsible for our actions because our actions are a result of how one is, mentally speaking, and it makes no sense to say that we can choose how we are. There have been many replies to this argument, one is that freedom and moral responsibility comes in degrees and can grow over time, reflecting the fact that "the way one is, mentally speaking" is increasingly shaped by past choices. Also, some choices may reflect more freedom than others (O'Connor and Franklin, 2019).

In contrast to the logical arguments against free will there are also empirical arguments. Empirical arguments against freedom come mainly from fundamental physics and psychology broadly speaking, including neuroscience and biology.

From neuroscience and psychology we see results showing that we can believe that we have freely chosen to do something that was in fact artificially induced. Also people with certain neurological disorders sometimes seem to do some action on purpose, at the same time as they sincerely believe that they are not directing them. Benjamin Libet is famous for conducting some experiments that seemed to prove that there exists some "preparatory" brain activity shortly before a subject does something seemingly spontaneous. Libet thought that this means that the brain decides what to do before we consciously choose the action (O'Connor and Franklin, 2019). On the basis of these experiments one can argue that conscious choice is an illusion, because no matter what you chose to do, your brain has already chosen it before you are aware of making a choice.

Philosophers have also argued that we have good empirical reasons from fundamental physics, to believe that the world is causally determined. And since humans are part of the

physical world, our choices must also be causally determined. Many thought that Newton's theory that the world follows simple laws of motion proved this. However, the quantum revolution in the early twentieth century has made matters more difficult. The implications of quantum mechanics on the causal structure of reality is still not well understood, and there are competing nondeterministic and deterministic interpretations. It is possible that the indeterminacy we find on small-scale "cancels out" at larger scales of animals and people, so that behavior on these larger scales would still be completely deterministic. This general rule is somewhat challenged by a number of exceptions, for instance quantum effects in some biological processes, giving rise to a new field of quantum biology. Still, even with macro quantum effects, it is not clear how it could result in useful, mental freedom. In any case, current science does not *decisively* support the idea that everything is predetermined by past events completely out of our control (O'Connor and Franklin, 2019).

The assumption that nature is governed by fundamental, exceptionless laws usually goes unquestioned in the physical sciences. Laws of nature are implicitly thought of as causing everything that happens. If the laws governing our world are deterministic, then in principle everything that happens can be explained as following from states of the world at earlier times. Many theories of laws of nature hold that the laws are in some sense *necessary*. However a growing class of philosophers hold that (universal, exceptionless, true) laws of nature simply do not exist. Among those who claim this are influential philosophers such as Nancy Cartwright, Bas van Fraassen, and John Dupré. For these philosophers, determinism is simply a false doctrine. This does not mean that concerns about human free action are automatically resolved; instead, they must be understood in the light of whatever account of physical nature without laws is suggested (Hoefer, 2016).

I do not propose that natural laws do not exist, however I do agree that they might not be strictly necessary in all possible worlds. There might be some alternative universe where the laws of nature are completely different than in our universe. It also seems easier to explain free will if the laws of nature are not the only strictly necessary explainers of all that happens. In the case of free will, I think it is possible that mental powers can have causal influence in addition to the natural laws. However they are not mutually exclusive; stable laws and a physical reality that behave in a predictable way seems to be necessary in order for us to plan ahead and make free choices. This intuition that natural laws are not only compatible but

even necessary for free will, is one of the main motivations for “compatibilism”, which is the view I will discuss next.

Compatibilism

Compatibilism is the view that free will is possible within a causally determined universe. In the early modern period a two-step strategy for defending compatibilism emerged. Those philosophers who argued for this two-step strategy are known as “classical compatibilists”. The first step was to argue that the opposite of freedom is not determinism, but external limits that keep you from doing what you want to do. Hobbes for example, stated that freedom is the absence of all the things that limits your actions, which is not within your own nature. This idea led many compatibilists to develop an analysis of the “freedom to do otherwise” and “self-determination” that were based on the agents preference or desire. If an agent had the freedom to do something other than X, then she would have done it, if she prefered or wanted to. The “freedom to do otherwise” does not require that you are able to act in opposition to your strongest motivations or desires, just that if you had desired something else more strongly, than you would have done that instead. An agent “self-determines” her actions, if they are caused by her strongest desires or preferences at the time of the action (O’Connor and Franklin, 2019).

The second step of classical compatibilism was to argue that it is impossible to analyze free will in a deeper, more robust sense of freedom because this will lead to many difficulties. Immanuel Kant, Thomas Reid and C. A. Campbell are some philosophers who have tried to capture a deeper sense of freedom. These philosophers argued that the classical compatibilists analyses of the “freedom to do otherwise” and “self-determination”, are not enough for free will and maybe also incompatible with it. They argued that the freedom to do otherwise is not enough, free will means not only that an agent could have *acted* differently if she had willed to, but also that she could have *willed* differently. Free will is more than free action, they argued that self-determination requires that the agent herself, not her desires and preferences, causes her free choices and actions (O’Connor and Franklin, 2019).

Against these claims, the classical compatibilists argued that while it is intelligible to ask whether a person willed to do what she did, it is incoherent to ask whether a person willed to *will* what she did. In response to the libertarians claim that the agent, and not her desires, are needed for self-determination, they objected that this removes the agent from the natural causal order, which is impossible for humans. An implication of this is that free will is not only compatible with determinism, but even requires it. This was a commonly shared assumption among compatibilists in the mid twentieth century (O'Connor and Franklin, 2019).

There are two features of free will that are most discussed: The "freedom to do otherwise" as discussed and "sourcehood" (that the agent is the source of the action). Most philosophers hold that free will includes both these aspects. The most common feature seems to be "The freedom to do otherwise". It seems that the freedom to do otherwise must consist in something more than just a mere possibility of something else happening. It is more plausible that it is an ability or power of the agent herself. This is the addition in the concept of sourcehood. A satisfactory explanation of the freedom to do otherwise, must give an account of what kind of ability this might be, and why this kind of ability gives us freedom. If determinism is true, then all our actions are consequences of the laws of nature and past events. With this framework we have no control over either the past or the future, and since we have no power to influence the laws of nature, the consequences of these things, including our present actions are not up to us. Some have argued that compatibilism requires the freedom to do otherwise to be some kind of ability to break a law of nature, and it does not seem probable that humans are capable of that (O'Connor and Franklin, 2019). Next we will look at some contemporary accounts of determinism that might be able to give a better account of the freedom to do otherwise.

Contemporary compatibilism

The Consequence Argument implies that determinism and the freedom to do otherwise is incompatible. Assuming that determinism is true, it states that:

1. "*No one has power over the facts of the past and the laws of nature.*
2. *No one has power over the fact that the facts of the past and the laws of nature entail every fact of the future (i.e., determinism is true).*

3. *Therefore, no one has power over the facts of the future*" (McKenna and Coates, 2020).

Several contemporary compatibilist theories attempt to explain the freedom to do otherwise in a way that is compatible with causal determinism. Some have argued against the first premise by trying to show that the way we act can change the course of time (McKenna and Coates, 2020). However this would break with determinism, because according to determinism all the past and the future is fixed and no free agent can act in a way that changes what is already predetermined to happen.

Other compatibilists have argued against the first premise by trying to show that a person in fact can act in such a way that a law of nature would not obtain. Not in the sense that the agent breaks a law of nature, but that this law of nature would not apply to her action (McKenna and Coates, 2020). This also seems to break with determinism because according to determinism everything must necessarily follow the laws of nature at all times and in all circumstances.

Michael Slote attempted to refute the Consequence Argument by showing that its central inference is invalid. He pointed out that notions like "unavoidability" are sensitive to contexts. Even though there are many unavoidable facts about the laws of nature and previous events, this kind of unavoidability is misapplied when it concerns aspects of a person's agency. It is claimed that these facts are unavoidable for a person, but from this a conclusion is drawn that the very actions a person performs are unavoidable for her. Yet this, Slote and other compatibilists suggest, is to draw illegitimate incompatibilist conclusions from reasonable claims about unavoidability (McKenna and Coates, 2020). I agree that the consequence argument as it is stated above, should not be seen as an irrefutable argument against compatibilism, after all the whole point of compatibilism is that free will is not in conflict with determinism, and even requires determinism. For compatibilists, freedom should not involve any hint of the ability to change the future or break any laws of nature, instead compatibilists must give an account of freedom that naturally arises from the deterministic way the world is.

I have much sympathy for the compatibilist sentiment, it attempts to unite the two strong intuitions that both determinism is true and that we have free will. However, even if compatibilists are able to make sense of a kind of freedom that agrees with determinism, I

think a deeper sense of freedom will require the ability to affect future happenings. This means giving up the part of determinism that claims that all of time is fixed for all eternity. It seems that this deeper sense of freedom can only be captured within a libertarian framework. I will give a more thorough account of Libertarianism in the next chapter, since this is the view I argue for in this thesis.

Chapter 2: Libertarianism

Libertarianism is the incompatibilist position that aims to refute determinism in order to keep the possibility of free will. The two main claims for the Libertarian view is that 1) Free will is possible, and 2) Determinism is false.

John Duns Scotus was the first to defend a strong libertarianism in the medieval period. He claimed that by its very nature, will has to be the cause of its own activity (O'Connor and Franklin, 2019). This idea expresses the heart of libertarianism, that free action must be caused directly by the agent. In recent years philosophers have considered the experience of agency, and there is a large discussion about what it consists in, and if it might support an indeterministic theory of free action. It is sometimes claimed that our belief in our own free will is epistemically basic and reasonable, and therefore need not be proved independently. Most philosophers hold that some beliefs have that status, on the threat that we might not have any justified beliefs at all. It is controversial which beliefs are allowed under this category because it is controversial which criteria a belief must meet, in order to be qualified for that privileged status. It might be necessary that a basic belief is "instinctive" for all or most humans, that it is part of normal experience, and that it is central to how we understand the world. Our belief in free will do seem to meet these criteria, but it is debated if they are sufficient (O'Connor and Franklin, 2019).

There are three main libertarian options for understanding sourcehood or self-determination:

1. Non-Causal Libertarianism
2. Event-Causal Libertarianism
3. Agent-Causal Libertarianism

Non-causal libertarianism argues that the power of self-determination does not need to be caused, we can control our choice, simply because it is ours and happens in us. There is no special kind of causation that makes it happen, it is an intrinsically active event. There might be causal influences on our choice, but this is not necessary and those influences are completely irrelevant to understanding how the choice happens. Since our choice is not completely determined by previous events, it is free and under our control simply because it

is ours. This view does not have wide support among libertarians because self-determination seems to be an essentially causal notion (O'Connor and Franklin, 2019).

Therefore I will focus on the two other forms of libertarianism, namely Event Causal Libertarianism and Agent Causal Libertarianism. Ultimately I will argue that Agent causality is the most promising version if we want to secure a more robust sense of freedom where the agent herself is the cause of her actions.

Event causality vs. agent causality

Most libertarians argue for an event-causal or agent-causal explanation of sourcehood. Both these accounts state that self-determination partly consists in the agent causing her own action, but they disagree on what the agent consists in (O'Connor and Franklin, 2019).

Event-causal libertarianism thinks that the agent causing her action is completely reducible to mental states, where the agent always causes her actions through these mental events. These mental events are the reasons for her actions (O'Connor and Franklin, 2019).

Agent-causal libertarians thinks that this event-causal picture fails to capture self-determination. They insist that self-determination cannot reduce to non-deterministic causation by appropriate mental states; in other words, agent causation does not reduce to neural events happening in the brain. On the other hand, many have argued that agent-causal libertarianism is incoherent, because the very idea of causation by agents that is not reducible to causation by mental states, is incoherent (O'Connor and Franklin, 2019).

The luck objection against event causality

The main problem for event-causal libertarianism is that it cannot give the agent any responsible control over her actions. This is also called the "luck objection", because in a situation where competing reasons suggest different actions, the choice made by agent-involving events would leave open to luck which decision would occur. The agent would have no further causal role in determining the decision. With the causal role of the antecedent

event already given, whether the decision occurs is not settled by any causal factors involving the agent. Since control is plausibly a causal matter, this fact provides a strong reason to conclude that the agent lacks the control required to be morally responsible for the decision (Pereboom, 2004, p. 275-276).

We could object that the agent still can "make up her mind", once everything that is causally relevant to whether the decision will occur has happened. However the luck objection can be restated at this point. If the antecedent events are all in place, and it is still unsettled if a particular decision will be made, the agent has no further role in determining whether it does or not, then she will still not be responsible for the decision. Her involvement would be restricted to being the subject of the decision, but this is not a causal role and the kind of control we search for must have causal power (Pereboom, 2004, p. 276).

While event causationists appear to be willing to reduce free will to purely physical processes happening in the brain, agent causationists insist that free will must come directly from the agent herself, there is nothing other than the agent that can be the cause of her actions.

The agent-causalist proposes to reintroduce the agent as a cause, not only as involved in events but more fundamentally, as a substance. This appeals to the controversial notion of substance-causation. This proposal is that the lack of sufficient control in the event-causal libertarian view, is supplied by a power of the agent-as-substance to cause a decision, and thereby settle which of the several possible decisions will occur (Pereboom, 2004, p. 278). I think this agent-causal picture of free will is the most promising, however there have been many objections raised against this position. In the next section I will present two of these objections and the replies to them.

Two arguments against agent causation

The luck objection again

It appeared first that agent causation might be able to solve the problem of luck, however the objection has been raised that in fact it does not really help. Consider some person called Jane. At a certain moment she, as an agent, causes a decision to accept a job offer. Until she

accepts, there remains a possibility that she will, at the last moment, choose to refuse the offer instead. Therefore in another possible world identical to the actual world, she could have chosen to refuse the offer in that moment instead of accepting it. There is nothing about the world prior to Jane's decision that can account for why she causes one decision instead of the other. It seems that this difference is just a matter of luck and therefore Jan cannot be responsible for her decision (Clarke and Capes, 2017).

However, if in fact Jane causing her decision means that she is exercising free will, then the difference between her causing a decision to accept, and her causing a decision to refuse, is not merely a matter of luck; it is a matter of how Jane exercises her free will. But how can we defend the claim that an agent's causing a decision is the same as exercising her free will? When one exercises free will, it is up to oneself whether one does one thing or another. A vast number of alternatives might be open, and the agent herself decides which alternative to choose. When she does choose, she is an ultimate source of her action. An agent-causal explanation seems to fit nicely into this familiar sourcehood conception of free will. If we assume that incompatibilism is correct, the explanation needs indeterminism to secure the openness of alternatives. And its requirement of agent-causation secures that the agent herself determines which alternative she choses, as well as being the origin of her action. If this explanation fits the sourcehood conception of free will, then it may be claimed that the difference in question between the two worlds is a matter of Jane exercising her free will differently (Clarke and Capes, 2017). I think it is reasonable to claim that the difference between the two worlds happens because Jane exercises her free will differently. It is not just a matter of luck whether she chooses to refuse or accept the job offer, it is a free choice on her part that causes the difference between the two possible worlds. This objection therefore is not a great problem for agent-causation in my opinion.

Substance causation

In cases of normal causation, for instance, when gravity causes things to fall, or the sun causes sunburns, it is widely agreed that causation by such things is reducible to the things they consist in. This is precisely what is denied when it comes to agent-causation. This denial raises the question whether any reasonable account of agent-causation can be given. Even some supporters of agent-causality think that this is doubtful and declares that agent-causation must be strange or even mysterious (Clarke and Capes, 2017). However, in general

it seems that all proposed solutions to the free will debate discussed so far, in some way or another are both strange and mysterious.

In this case, the strangeness may be viewed in light of the widely held presumption that, free will aside, causation throughout nature is, fundamentally, causation by events or states. Until recently, even most proponents of agent-causal theories have accepted this view. Therefore an appeal to agent-causation in a theory of free will strike many as highly implausible (Clarke and Capes, 2017).

However, in the last couple of decades, a growing number of philosophers have argued that causation by substances is ontologically fundamental. Some hold that, fundamentally, *all* causation is substance causation. Others advance causal pluralism, where many things like substances, events, properties, features, aspects, and facts can cause things, and causation by each of these kinds of things is equally fundamental. The pluralist picture is not one of competition but of interdependence. (Clarke and Capes, 2017).

Often such views are advanced on grounds that are entirely independent of free will, having to do with the nature of causation and causal powers generally. Substance causation that is ontologically fundamental is held to be pervasive, constituting the activity of substances animate and inanimate, macro and micro. Rejection of the view that causation by events or states is uniquely fundamental, often stems from a turn away from of a broadly Humean account of causation toward a neo-Aristotelian view, one that takes causal powers to be irreducible features of the world (Clarke and Capes, 2017).

In the context of a view of this kind, an appeal to causation by agents that is ontologically fundamental is no claim of metaphysical exceptionalism. Further, on a causal pluralist view, an agent-causal theorist can give a good explanation of what it means to act for reasons. It may be said that a free action is caused by the agent *and* caused by certain states of the agent, with causation by each of these things interdependent and equally fundamental. The relative strength of this view seems to depend on its general account of causation (Clarke and Capes, 2017).

In the last chapter I will suggest a minimal kind of agent that can act in virtue of her consciousness. I will argue that phenomenal consciousness itself has causal power. Now we will look at the problem with determinism and see if an agent-causal theory of free will can deal with determinism without completely rejecting all of its intuitive power.

Is determinism true?

One of the main claims of Agent-Causal Libertarianism which is the view I advocate in this thesis, is that determinism is partly false. However is there good evidence for this claim? Is there room for some sort of indeterminism in reality? Meaning that some events are entirely nondeterministically caused by agents and not events. This indeterminism must be present when free agents cause decisions and other free actions. What evidence do we have for these claims? (Clarke and Capes, 2017).

It is sometimes claimed that our experience when we make decisions and actions, establish good evidence that there is indeterminism of the required sort in the required place. Many find it incredible that how things seem to us when we act, gives us insight into the laws of nature (Clarke and Capes, 2017). I do not think this is incredible however, after all we are part of nature and the way things seem to be to us can be a legitimate part of trying to make sense of reality.

The scientific evidence for quantum mechanics is sometimes said to show that determinism is false. Quantum theory is indeed very well confirmed. However, there is nothing approaching a consensus on how to interpret it, and it seems like there has to be indeterminism of a specific sort at specific places in certain brain processes for free will to be possible. Indeterministic as well as deterministic interpretations have been developed, but it is not clear whether any of the existing interpretations are correct. Perhaps the best that can be said here is that there is currently no solid evidence that determinism is true (Clarke and Capes, 2017). I think that this lack of solid evidence creates a reasonable possibility for free will. I also think it is reasonable to look for this possibility in the realm of quantum physics. We will revisit this possibility later in chapter four of this thesis.

We see that there is much uncertainty about determinism and it seems that we really do not yet know whether determinism is true or not. It is important to understand what we mean when we argue for or against the truth of determinism. I do not argue that there are no laws of nature, as there obviously are some laws or at least deep regularities determining the shape and flow of our universe. These laws are not necessarily just a hindrance for free will, because they create a physical reality that we can understand and predict, and this predictable

reality makes it possible for us to make reasonable choices as agents. The problem with determinism in relation to free will seems not to be the existence of natural laws, but the claim that they are absolutely immutable and that they necessarily must be the only cause and explanation of everything that happens in our universe on all levels, in all past and future.

I think this idea of time being fixed and static without opening up to more than one possibility at any point, is the most problematic aspect in regard to free will. Free agents must exist in a space of possibilities with the ability to choose. To put it pointedly, it seems very strange to insist that my current choice between tea and coffee was already fixed as the big bang started. This would mean that all that needs to be explained is the original creation of the universe in just that way, and that any other seemingly creative act, since the big bang, such as our seeming free choices, cannot be original in any way, as it is all merely derivative.

The problem of dualism and agent causation

I want to defend a kind of non-reductive physicalist position in this thesis. However, right at the outset I want to point out that Jaegwon Kim gives a good argument against this position. He argues that if mental events are distinct from, yet wholly "realized by" physical processes, there can be no causal factors beyond the physical. This argument assumes physical causal closure and that mental causes do not systematically overdetermine events caused by physical factors. If this argument works it apparently forces us to accept either a complete identity between mental events and specific physical events, or dualism between the physical and mental (O,Connor, 2005, p: 338).

Does this mean that the agent causalist must be a dualist? Agent causal power cannot be reduced to underlying physical processes because then it is no longer an ontologically irreducible power. It may be enough to suppose that agent causal power and its properties are ontologically emergent, while still being powers and properties of the physical organism. This is a strong form of property dualism, which states that the mental and physical are fundamentally the same substance, but with different properties (O,Connor, 2005, p: 342).

I agree with Timothy O'Connor that a framework of property dualism and emergence of the mental from the physical might be the most promising option to secure a possibility for free will. This requires a metaphysical understanding of emergence. The problem of dualism

between the mental and the physical and the idea of emergence will be the main topics in the next chapter.

Chapter 3: Mental causation

Is the mental reducible to the physical? If so, then is mental causation weakly or strongly emergent from physical causation? These are the main questions for this chapter. As to the first question, my answer in this thesis is ultimately no, the mental is not reducible to the physical, however this need not imply that it is not in a broad sense constituted by the physical. With regards to the second question I will argue that the mental must be strongly emergent, since weak emergence does not seem able to give us the kind of remarkable "downward" causation that the mental apparently manifests. We will begin at the beginning by first asking the question of whether something mental exists at all?

Do the mental exist?

In order to exert our will, it seems that we must have a mental power that is able to direct or choose our actions. For free will to be possible, it seems clear that something mental must exist and that it has a real causal power in the physical world. According to the "Eleatic Principle" *power is the mark of being*, meaning that for something to exist it must have real causal power (Armstrong, 1978 and Oddie, 1982 in Robb and Heil 2019). If this principle is true, which we assume here, then if something mental exists at all, it is reasonable to think that it has causal power, able to affect the physical world. However there are many problems with mental causation which we will explore further in this chapter.

Descartes famously pointed out that minds and bodies appear to be two different kinds of substances. Bodies are extended in space and have no experiences or thoughts. Minds on the other hand are unextended, thinking and experiencing "souls". The mental seems to be something very different from the physical, and there is much disagreement on how to build a comprehensive picture of agents possessing both mind and body (Robinson, 2017). It seems plausible to assume that either the mental is ontologically distinct from the physical and cannot be reduced to the physical, or the mental and physical are not distinct and have the same kind of causation.

The problem of interaction

Descartes accepted the intuitive belief that mind and body causally interact. It seems obvious that thoughts and feelings can move the body and feel what happens to it. However, if mind and matter are two completely different substances, it is hard to understand how they can interact. If something mental is to cause a body to move, it must somehow be in contact with it, but there are many seeming problems with this, for instance, since the mental has no spatial location, how could it come into contact with a purely physical body? (Robb and Heil 2019). We have good reasons to think that the mind is constituted by the brain, however science has not found any neural activity that can be said to directly cause or be caused by the conscious mind. This problem is often called the "mind-body" problem in philosophy.

There is a thesis about causation which holds that any causal relation between two things, depends on a *nexus,* or *common interface* where cause and effect are connected. This interface need not be spatial contact, but it seems that if we are to make sense of mental causes in a physical world, such an interface must be found. However it might be that this principle rests on an outdated conception of causality, holding no place in modern physics (Robb and Heil 2019). In the next chapter I suggest a kind of strong emergence called "fusion emergence" which might be interpreted as a kind of common interface between the mental and the physical. First I will present epiphenomenalism, in this theory the mental and physical need not have any common interface because the mental does not really exist in its own right, it is only a kind of "shadow" cast by the physical.

Physical causal closure and epiphenomenalism

How can mental causation be possible in a world that is physically closed? The causal closure principle demands that every physical event must have a wholly sufficient physical cause. (Walter and Heckmann, 2003, p. 141).

Even if every physical effect has a sufficient physical cause, we might think that maybe some physical effects can have a mental cause in addition to a sufficient physical cause. However,

there is wide support in the litterature for the principle that we can have no systematic overdetermination of physical effects. The dualist's options are apparently limited to either embracing parallelism, which states that bodies and souls as running in tandem, with no causal influence in either direction, or embrace epiphenomenalism (Robb and Heil 2019).

Epiphenomenalism is the view that mental events are caused by physical events in the brain, but have no effects upon any physical events in turn. Behavior is caused by muscles that contract upon receiving neural impulses, and neural impulses are generated by input from other neurons or from sense organs. On the epiphenomenalist view, mental events play no causal role in this process (Robb and Heil 2019). The metaphor often used is that the mind is just a shadow cast by the brain.

Many philosophers recognize a distinction between two kinds of mental events. Namely qualia or phenomenal experiences like feeling pain or tasting coffee, and propositional attitudes like beliefs and desires. Arguments about epiphenomenalism may concern just one or both these types of mental events. The two types are often connected, however, through beliefs that one has one's experiences. So if one claims that pains have no physical effects, then one must say either that pains do not cause beliefs that one is in pain, or that beliefs that one is in pain are epiphenomenal. For, if pains caused beliefs that one is in pain, and the latter had physical effects, then pains would, after all, have indirect effects in the physical world. But epiphenomenalism claims that mental events have *no* effects in the physical world at all (Robinson, 2019).

Libet's studies

It is often argued that empirical evidence overwhelmingly supports epiphenomenalism. These arguments often refer to the experiments done by Benjamin Libet, that were mentioned in the paragraph about Determinism. These experiments apparently prove that conscious willing must be a product of non-conscious processes that do the real causal work. When self-awareness is present, the experiments have shown that it occurs too late to be the cause of the relevant actions rather than their result (Libet 1985 in Van Gulick, 2018). Self-awareness according to these arguments turns out to be a psychological after-effect rather than an initiating cause (Van Gulick, 2018). This is problematic for a position on free will

conditioned on being aware or self-aware. If the brain has already chosen an action before the conscious part of the brain is even aware of it, it does indeed look dark for the kind of free will we are supposing here.

Objections to Libet's studies

The implications drawn from these experiments have been criticized on several grounds. In some cases, there are technical criticisms of the methods or statistical analyses used. Two main criticisms are offered. First, many of the experimental conditions involve meaningless setups or meaningless decisions that bear little resemblance to real-life decisions, where people act with reasons and personal preferences. These oddities and simplifications may make psychological mechanisms ineffective that would normally be active in more realistic cases and that would prevent illusions about the relation between our conscious reasons and our behavior. Second, the fact that there are some cases where unconscious influences affect our behavior does not show that we are never, or even not usually, acting in a way that would be rational, given our particular long standing beliefs and preferences (Robinson, 2019). We see that the Libet experiments might not be as trustworthy as is often supposed, and even if they are proven to hold, this does not necessarily put the nail in the coffin for free will.

Objections to epiphenomenalism

There are several objections to epiphenomenalism, however I will briefly give two here. First, the absurdity objection. Second, the self-stultification objection. We respectively go through each briefly here.

The absurdity objection

The absurdity objection states that epiphenomenalism is absurd. It is plain obvious that our pains, thoughts, and feelings make a difference to our behavior. It seems absurd to insist that all our behaviour would be just the same if we had no sense impressions, no pain, pleasure, thoughts or feelings. Epiphenomenalists can reply that it is never obvious what causes what,

however given its strong intuitive absurdity one may ask that they provide very convincing arguments in order to show that the mental cannot cause anything (Robinson, 2019).

The self-stultification objection

Epiphenomenalism implies that we have no actionable knowledge of our own minds — and thus, incompatible with us even knowing in practise the idea that epiphenomenalism is true. One variant has it that we cannot even succeed in referring to our own minds if epiphenomenalism is true. Another problematic aspect is that anything an epiphenomenalist utters or writes about epiphenomenalism, in physical practise cannot be caused by what is epiphenomenal, namely the mind or consciousness of that epiphenomenalist.

If these claims are right then epiphenomenalism is, at the very least, caught in a practical contradiction, in which they must claim to know, or at least believe, a view which implies that they can have no reason to believe it (Robinson, 2019). Even though epiphenomenalism perhaps can save a weak form of mental causation, while staying true to physical causal closure, the objections against this position are so strong that we must find other options. One option is property dualism, which we will explore next.

Property dualism

Property dualism holds that mental properties are dependent on, but not necessarily reducible to physical properties. Some property dualists give this status only to the class of mental property called “qualia”, meaning the “what it is like” features of conscious experience. Other property dualists are willing to extend the thesis to all mental properties.

Suppose that a robust form of property dualism is true. Can mental substances or events have causal power in virtue of their mental properties alone? According to Robb and Hail (2019), the arguments against soul–body interaction, can be re-entered here in terms of properties. They argue that if we cannot explain the connection between soul and body, it is just as hard to explain how non-physical properties can have any impact on the physical world. Also, the principle about closure, namely that every physical effect must have a sufficient physical cause, does not seem to be satisfied by holding mental properties to be a sufficient physical

cause. If we state that a sufficient physical cause should be sufficient in virtue of its physical properties alone, the efficacy of mental properties are again threatened (Robb and Heil 2019).

A similar option is to argue for non-reductive physicalism, this view agrees with property dualism that mental properties are not physical, but separates from property dualism in that the mental depends on the physical. Mental properties are “realized” or “constituted” by physical properties (Robb and Heil 2019). The discussion above is largely dependent on the idea that “the physical” consists of fundamental particles and their interactions. I will argue later that non-reductive physicalists can still hold mental properties to be physical, if we allow for strongly emergent properties to count as physical. We will explore this option in the next chapter and consider the possibility that mental properties need not necessarily be realized or constituted by fundamental particles, in order to be counted as physical.

Physicalism

Physicalism is the most successful ontology to this day, and has outcompeted all dualistic and idealistic competitors. It is reasonable to believe that everything that exists, including mental properties, is composed of fundamental physical properties. Three questions then arise; What is a physical property? What is a fundamental physical property? and what does it mean to say that something is composed of fundamental physical properties? Physicalism should not rule out the existence of mental properties, unless you're an elimitavist and claim that mental properties do not really exist. If the physical is defined by contemporary physics as it stands, then strictly speaking physicalism is probably false, because contemporary physics is most likely still incomplete; future discoveries, like the rapid developments we see in quantum physics, should also count as physical (Walter and Heckmann, 2003, p. 3).

Various formulations of physicalism have been suggested. Frank Jackson has given an influential formulation of physicalism which states physicalism as the supervenience thesis: "*Any world which is a minimal physical duplicate of our world is a duplicate simpliciter of our world*" (Jackson in Loewer 2011, p. 197). A physical duplicate, duplicates both the physical facts and the laws of physics. The idea is that once the physical facts and laws of our world are fixed, then all other facts about our world are also fixed (Loewer, 2011, p. 197).

This formulation expresses the core physicalist idea that the world is nothing over and above the physical.

The micro-macro distinction

It is difficult to understand how mental states like being in pain or tasting coffee can be reducible, for example to some sort of neurological state in the brain, which again can be reduced to fundamental particles. Mental properties like pain, are often explained as "higher-level" properties, that come about by appropriate "lower-level" properties, which are the physical "realizers", or fundamental building blocks of any particular pain or smell or any other mental experience (Robb and Heil 2019). In this sense everything we experience through our senses are high-level properties. This distinction between high and low-level properties is also called the micro-macro distinction, and it is generally given as a rough, intuitive distinction. Micro properties like atoms and molecules exist on the lowest levels while macro properties like houses, trees and the taste of coffee exist on the highest levels.

This distinction between micro and macro properties, gives us a layered model of the world, where everything is organized in a hierarchical structure of levels and the bottom level consists of elementary particles. As we go up to higher levels we find atoms, molecules, cells, larger living organisms, and so on until we have the human brain. At each level we find new properties, activities and functions that we did not find at the previous level (Kim, 1998, p.16). This layered model has inspired much of the philosophical debates about reduction and reductionism, the mind body problem, emergence, the status of the special sciences and the possibility of a unified science (Kim, 1998, p.16).

The crucial question is how the different levels relate to each other. How is consciousness related to biological and physical processes? One might argue that properties at every level higher than the bottom level, are reducible to lower-level properties and ultimately to the fundamental properties of physics. Or one can take an anti-reductionist stand and argue that some phenomena, and consciousness in particular, cannot be reduced to its underlying levels (Kim, 1998, p.17). I will take this anti-reductionist stand in this thesis, as it seems to me that something essential to a high level property, for example the taste of coffee, is taken away if it can be reduced to the properties of its constituent parts and their relatively straightforward

combination. Before we dive into this, I will first give an account of reductive physicalism, as well as the exclusion argument, and then I explain non-reductive physicalism in more detail.

Reductive physicalism

Reductive physicalism is the standard version of physicalism. It states that the nature of reality is nothing but spatiotemporal arrangements of fundamental physical objects and properties. Different arrangements of these elementary objects and their properties, account for all the vast variety we find in the world. Humphreys (2016) calls this the *generative atomism* version of physicalism: In this sense, "*The fundamental laws of the universe govern the spatiotemporal arrangements at the lowest levels, and everything else that goes on is determined by those arrangements*" (Humphreys, 2016, p.8). Generative atomism inspired the physicalist idea that everything is built up by micro-physical properties. This is similar to the definition of physicalism given by Jackson above, but adds that everything that exists is reducible to the micro-physical.

Generative atomism in this basic form has both a synthetic and an analytic component. "*The synthetic component says (1) that there is a collection of elementary entities from which all other legitimate objects in the domain are constructed, (2) there is a fixed set of rules that govern the construction process, and (3) as a consequence of (1) and (2), all entities are either atoms or are composed of atoms. The analytic component asserts that any non-atomic object can be uniquely decomposed into its atomic components using an explicitly formulated set of decomposition rules*" (Humphreys, 2016, p.13). With atoms in this context, Humphreys means basic units of matter, the word "atom" means "indivisible" (Humphreys, 2016, p.23).

Generative atomism seems to rule out all non-reductive or emergent properties, to the extent that we take the fundamental elements as those proposed by fundamental physics to be the only existing objects. Any emergent property, like a mental property M, could always be uniquely decomposed into atomic components P that explain the causality of M. However, this "exclusion" of M by P, as we will see below in the section on the exclusion argument, can be answered in a number of ways. Some of these turn out to justify only weakly non-reductive ("M"-like) properties, whereas some are more ambitious in their response.

It is also worth mentioning that generative atomism has a long history, and seems to trace back at least to Newton's idea that elementary physical particles are the fundamental building blocks of the world. Newton's picture of a world built out of fundamental particles is still vastly influential. Sometimes reductive physicalism is also simply called fundamental physicalism, because this view is so closely related to, and usually committed to the existence of the entities and properties of fundamental physics, such that everything else that exists is determined by those entities and properties in virtue of some determination relation. If the realm of fundamental physics is causally closed, then it is reasonable to assume that all the events that causally determine some property must occur in the domain of fundamental physics. Any events outside that domain seem to be causally redundant. This is why reductive physicalism in its most extreme form, completely rules out everything that is not fundamental. This idea is typically indicated by the word "exclusion", because it implies that the ontology of fundamental physics, excludes all other causally relevant ontologies.

The Exclusion Argument

The exclusion argument, or exclusion problem, is the idea that if the realm of fundamental physics is causally closed, there is no causal room for entities distinct from those included by fundamental physics. The argument can take my forms, but supposing for instance that a causally efficacious mental property M is instantiated, then - supposing fundamental physics is causally closed - it will be realized by some particular physical property P. Now, we can in principle explain all causal effects of M by appealing to P, so why include M in our ontology?

In this case, P seems to "exclude" M, or make M irrelevant, seemingly leading M to become epiphenomenal. In general any property, B for biological, C for chemical, etc. will be "excluded" by P, and might therefore seem dispensable (Humphreys, 2016, p.72).

This argument reflects the intuitive notion that, when the properties on the fundamental micro-level of a physical system are fixed, so are the properties of all its macro-levels. This relation is usually taken to imply that the micro-mechanisms do all of the causal work, so that the micro-level is causally complete (Hoel et al, 2013, p. 19790). The exclusion argument raises the perplexing question for us: How can high-level mental properties, like for instance

the feeling of *pain*, have any causal power if there is only causal power on the fundamental physical micro-levels? (Robb and Heil 2019).

There are several ways of answering the "exclusion" problem, and the main classificatory scheme divides the proposed solutions into three categories, as either autonomy, inheritance or identity solutions. I discuss each of them in turn below.

Autonomy solutions

Philosophers who go for the autonomy solutions argue that the mental properties have their own autonomous casual power and that this is not excluded by the causal power of physical properties. One version of this solution argues that psychological explanations and other explanations in the special sciences are independent of physical explanations. We can say that psychological explanations exist on a higher level of abstraction then their lower-level physical realisers, and appeal to their own distinct kinds and laws.

In this way, explanations given by the special sciences work independently of the explanations given in the lower-level physics, which means that mental and physical causes can coexist (Robb and Heil 2019). One worry with this approach may be the extent to which "explanations" are detached from any objective ontology; some philosophers may imply here a relativistic worldview that is not shared by everyone. On the other hand, to the extent that the explanations are anchored in an objective causal order, then it is not clear how the framework of mere explanations can offer us a satisfactory ontological resolution.

The exclusion problem entails that a mental property and its micro-physical realizers compete for causal efficacy over the same effect. However, another approach may be to say that for instance a mental property is not necessarily threatened by exclusion if it is relevant to a different property of the behaviour than the physical realiser. The mental property and the physical realizer need not be in competition with each other, because they have separate autonomous causal lines to different properties of the effect. Behavioural properties appear to be multiply realizable, there can be many different physical realizations of the same behaviour. The physical realization is only relevant to the particular *way* the behaviour is performed. It seems natural to suppose that the higher lever property of the behaviour itself, regardless of how it might be physically realized, is a mental property (Robb and Heil 2019).

The autonomy solutions secures a causal role for mental properties without conflicting with the completeness of the physical. However it cannot straightforwardly avoid overdetermination. Autonomy solutions present us with mental and physical properties, where each should be a sufficient cause for the behavioural effect. Even though the physical and mental property is responsible for different properties of the effect, it might be hard to avoid overdetermination, because the effects of the behavioral property are produced twice. Both by the mental cause and the physical realizer (Robb and Heil 2019). I think that it is possible to carve out a stronger autonomy solution to the exclusion problem that can avoid overdetermination. This is done through the process of fusion emergence, where the new mental property gets causal power in its own right because it's micro-physical constituents disappear or get "smeared out" in the fusion process. This sounds very strange, however it looks like this kind of process actually happens in the realm of quantum physics. I will present this strong autonomy possibility in the next chapter. First we will look at two other solutions to the exclusion problem and present the idea of emergence.

Inheritance solutions

Autonomy solutions make it look like the causal powers of mental properties are independent of their physical realizers, however unless one is given an explicit demonstration of how this is supposed to work, one is left with a suspicion of parallel operations that ultimately still overdetermine their effects (Robb and Heil 2019).

A different approach to seeking autonomy for the mental is to try to leverage the causal power in the micro-physical in order to achieve the non-reductive effects on higher levels. The idea is that the non-reductive properties will "inherit" their causal powers. Jaegwon Kim, for instance, has attempted to bind the mental powers more closely to their physical realizers in this way. This means that the mental properties are so closely related to their physical realizers that they "inherit" their causal powers. The mental does not compete with the physical, but cooperates with it. We also avoid overdetermination since the mental works through the physical. Some versions of the inheritance solution claim that the high-level mental property obtains some weaker or "lower-grade" form of causal relevance from its physical realizer. There is a distinction between a robust form of causal efficacy on the lower,

physical level and a weaker causal efficacy on the higher mental level. This view gives a derived form of relevance to mental properties, in a way that respects the completeness of the physical as well as the principle of no overdetermination (Robb and Heil 2019).

We can worry that this weakening of the mental results in epiphenomenalism, however we can look for an inheritance solution where mental properties are causally efficacious in the same sense as their physical realizers are. But can this option avoid overdetermination? We could say that a mental property is "immanent" in its physical realizer. In that case the mental property simply inherits the casual work done by the physical realiser. We avoid overdetermination because the work done by the mental property is included in the work done by the physical realizer. The physical realizer does not exclude, but includes the mental property (Robb and Heil 2019).

Identity solutions

The main problem with both autonomy and inheritance solutions is that both implies that mental and physical properties are numerically distinct, however intimately related they might be. Identity solutions on the other hand tries to bridge this gap. It claims that any mental property just *is* its physical realizer. Then there is no worry that one excludes the other, nor is there any mystery of how the mental property can work through the physical, because they are exactly the same. This excludes the possibility of the mental being multiply realisable, if the mental and the physical are the same thing, then a particular mental expression depends on a particular physical realizer, in order to be exactly the same in each new instance of its expression. This argument does not aim to show that the mental is distinct from its physical realizers, but that what looks like one kind of mental property actually is many. There are different kinds of pain, realized by many different physical properties, despite them all having the same name, "pain". There are similar mental properties, but they all have distinct physical structures (Robb and Heil 2019).

The price one must pay to adopt this solution, is to give up believing in pain as a single natural kind. But is it not possible to keep this belief by claiming that a property can be both mental and physical at the same time? The problem with this claim is that it appears to raise the exclusion problem again. According to the completeness of the physical and the principle

of no overdetermination, we are forced to say that the property is only causally efficacious as physical and not as mental (Robb and Heil 2019).

Some philosophers like Fodor, Baker and Shapiro, argue that exclusion might not necessarily threaten the efficacy of mental properties. We seem to have no problem with accepting for example biological or geological properties as causally relevant, even though they are not identical with their physical realizers. Then there is no reason to think that mental properties are different. Others turn this argument around and insist that the causal power of biological properties and other properties found in the "special sciences" should not be accepted as a matter of course (Robb and Heil 2019).

Jaegwon Kim has tried to solve this problem by including aggregates of basic particles as part of the physical domain (Kim, 1998, p. 113). However as a reductive physicalist he must be committed to the idea that all aggregated causal power ultimately is reducible to the causal power of the micro domain. We will look into Kim's account in further detail below in the section on weak emergence. Before we go into the concept of emergence however, it is worth clarifying a bit what we mean by non-reductive physicalism.

Non-reductive physicalism

Non-reductive physicalism grew out of "functionalism", the doctrine that mental properties are functional properties. For example, pain is being in a mental state caused by tissue damage which again causes responses like crying or efforts to heal the damage. Because this is describing pain at a functional level, it seems that pain can be "instantiated" or "implemented" not only in humans, but also in animals and perhaps ultimately in artificially intelligent systems in the far future. In general, it seems that a functionally isolated mental state, like pain or other experiences, can be realized by many different kinds of physical and maybe even non-physical systems (Robb and Heil 2019). This aspect of functionalism, which holds mental states like pain to be "multiply realisable" in many different kinds of physical systems, makes reduction difficult and is one of the main motivations for the non-reductive physicalist position.

The idea of non-reductive physicalism is therefore that although the fundamental ontology is physical, nevertheless, we cannot reduce and/or explain all things merely by using the properties, laws and concepts of fundamental physics. That mental properties like pain are constituted by physical properties seems fine, but that they are identical to physical properties seems highly implausible for the non-reductive physicalist. Again, the main problem with property-identity claims is that the same mental property apparently can be realized by multiple physical properties. When both Mary and John hold the same belief, for example that the earth is a planet, the physical realizers of Mary's belief can be vastly different from the physical realizers of John's belief. If a single mental property, like a belief, can be realised by several different physical realizers, psychophysical reduction seems impossible (Walter and Heckmann, 2003, p. 4).

Fodor, Putnam and others suggested that mental properties that are "types" cannot be reduced to physical properties, while mental properties that are "tokens" on the other hand are identical to physical properties. This new distinction however brings out further issues. The main problem for the non-reductive physicalist, is to explain how irreducible mental properties can exist in a world that is purely physical. Non-reductive physicalists have responded that even though mental properties are irreducible, they still depend on fundamental physical properties and this allows them to count as physical. For years it was thought unproblematic that mental properties were "realized" by physical properties in this way, thus securing both the irreducibility of mental properties and their broadly physical nature (Walter and Heckmann, 2003, p. 5).

However, in the nineties Jaegwon Kim argued that realization is necessarily a reductive relation. He introduced the "causal inheritance principle" discussed above, where realized properties inherit all their causal power from their realizers. All causal power found in realized properties should be reduced to their realizers. This raises the suspicion that realized properties are never causally relevant. If we accept the principle that all real things must have causal power, we are only one step away from eliminativism (Walter and Heckmann, 2003, p. 5-6).

Kim's latest view of the matter is that mental predicates and concepts might play a crucial role because we use them to group physical properties in ways that are essential for our descriptive, communicative and explanatory purposes, but seems to hold that real mental

properties with causal powers of their own is an idle dream (Walter and Heckmann, 2003, p. 5-6). This indicates that Kim's use of the word *explanatory* in this context, amounts to a relativistic turn with respect to the content under discussion, as it pertains to the explanatory practises of humans. The realism we need to invoke in the context of this thesis, however, with respect to consciousness and free will, is something that is not plausibly a matter of an explanatory practise, *unless* one is already an ontological eliminativist or reductionist. But we will return to this point later as we will explain Kim's casual account in more detail.

Ultimately I think that a kind of non-reductive physicalism must be true. I agree that mental properties cannot be reduced to something purely physical. However this depends on what we mean with "purely physical" and "reduction". I do think that the mental is ultimately physical, but as we will see in the next chapter, I argue that "the physical" in a broad sense must somehow also include strongly emergent, higher-level mental properties with independent causal powers. Before going into the next chapter, we will explore the notion of emergence and I will explain why I think mental causation must be strongly emergent.

Emergence

There are many ways to talk about things emerging from other things in general, but here I will frame the discussion in terms of it taking us from low-level physical realizers to high-level mental properties. Framing things in this way already presents a certain enigmatic quality to the picture, that some philosophers reject. However, by pursuing this path, I tentatively accept it for now, and resolve to find a solution within this framing of the problem.

I will start simply by separating between strong emergence and weak emergence. Strong emergence is when new high-level properties arise from the fundamental microphysical properties, but are not reducible, even in principle, to microphysical properties. Weak emergence is when high-level properties in some way or form unexpectedly arise from the microphysical domain, but are still in principle reducible to microphysical properties (Chalmers 2006, p. 1).

Weak emergence

Weak emergence allows that entities and features found in the special sciences are real, while also affirming physicalism, the thesis that all natural phenomena are wholly determined by fundamental physics, entailing that any fundamental-level physical effect has a purely fundamental physical cause. To deduce a weakly emergent, high-level property, one might need a large amount of calculation. However, when we examine the phenomena - if it is weakly emergent - it will turn out to be a straightforward consequence of low-level facts (Chalmers 2006, p. 1). Calculation is just one way to view it of course, the general gist is that it is “easy” or relatively straight forward (in principle, if not in practise) to figure out one from the other.

Special sciences describe structured phenomena and successfully predict their behavior through higher-level laws. Weak emergence explains these stable and distinctive phenomena as existing on high-levels and not on lower physical levels. In this view, there are molecules, cells, organisms, and conscious creatures, and they do in principle reduce to complex combinations of properties of basic physics, however they are distinct and explanatorily different from basic physics (O'Connor, 2020).

Problems with weak emergence

The weak emergentist allows that the ontology and dynamical laws of a complete physics, metaphysically determine all fundamental physical facts, and that it metaphysically determines all non-fundamental facts about the world. One might then explain all high-level phenomena and truths about concepts such as “cell”, “animal” or “pain” as referring to arrangements of underlying physical entities.

Such explanatory practises could simply be a way that we humans like to carve up the world for practical purposes. They need not necessarily have any special ontological import at the fundamental level, if one even believes in a “fundamental” level of realist ontology. This stance accepts that in practice we cannot give up the explanations given in the special sciences. However, should such considerations guide our views concerning the world’s ontology? (O'Connor, 2020).

One problem with having such considerations guide our fundamental ontology is that by listing all the premises in logical terms, we cannot have our cake and eat it too. If we talk about "higher-level" properties as applying to the *explanatory* level, we may admit of different things than when these "higher-level" properties are being applied to an *ontological* level. When we talk about "higher level properties" therefore, it is important to clarify if we are referring to an explanatory or or an ontological level. For clarity, I lay out the five premises that weak emergence generally accepts. These premises, as we will see, entail overdetermination, which leads many to reject one of the five premises.

Weak emergence generally accepts the following five premises:

1. *Supervenient Dependence*. Emergent features (properties, events, or states) synchronically depend on their base features such that the occurrence of an emergent feature at a particular time requires and is nomologically necessitated by the occurrence of a base feature at that time.
2. *Reality*. Emergent features are ontologically real.
3. *Efficacy*. Emergent features are causally efficacious.
4. *Distinctness*. Emergent features are distinct from their base features.
5. *Physical Causal Closure*. Every lower-level physical effect has a purely lower-level physical cause. (O'Connor, 2020).

These premises entail an unacceptable conclusion:

6. *Overdetermination*. Emergent effects are causally overdetermined by distinct, individually sufficient synchronic causes (O'Connor, 2020).

Thus, whether we conceive emergent causation as same-level or downward causal, the weak emergentist's commitments entail overdetermination. Finding such systematic overdetermination to be implausible, Jaegwon Kim concludes that we should reject premise 4, Distinctness, and embrace reductionism (O'Connor, 2020).

Alternative positions reject other premises. Eliminativists deny premise 2, Reality. Epiphenomenalists deny premise 3, Efficacy. Substance dualists and some strong emergentists deny premise 1, Supervenient Dependence. Strong emergence positions usually deny premise 5 on causal closure, and claim that emergent features are causally efficacious.

Kim's causal account of weak emergence

Jaegwon Kim presents an account of weak emergence, and argues that a weak kind of emergence is enough to explain mental causation.

Kim argues that strong emergence even takes away some of the essential meaning of "emergence", because the connection between the strongly emergent phenomena and its base properties is not a regular, determinative or necessitating relationship (Kim, 2006 p. 550). Kim wants to consider supervenience as an essential component of emergence, such that a concept of emergence will need to accept this proposition:

"Supervenience: *If property M emerges from properties N\,...,Nn, then M supervenes on N\,...,Nn. That is to say, systems that are alike in respect of basal conditions, N\,...,Nn must be alike in respect of their emergent properties*." ***(Kim, 2006 p. 550).***

Formulated as such, there is a strong determining relation from the lower to the higher levels, indicated by the "must" in this definition. Presumably, it would not follow as strongly in the other direction, because of the multiple realizability of emergent properties. In any case, this is meant to capture the gist of supervenience as generally construed, not a precise definition.

Logical, metaphysical and nomological supervenience

Another distinction with respect to supervenience, is whether the supervenience is logical, metaphysical or nomological. For Chalmers, for instance, the notion of strongly emergent phenomena means being systematically determined by low-level facts without being reducible to them, in the sense that "*they are naturally but not logically supervenient on low level facts*" (Chalmers 2006, p. 4). Thus, the sense of "being determined by" can either be logical, metaphysical or nomological, where logical means that it cannot be contradicted, metaphysical means roughly that it holds in all possible worlds, and nomological means that it is simply something that holds in this world by our particular laws of physics. Contemporary accounts of emergence states that weakly emergent or physically reducible features are taken to supervene with metaphysical necessity on their physical dependence base, while strongly emergent features on the other hand are taken to supervene with only nomological necessity (O'Connor, 2020). In this sense it looks like Kim assumes a stronger

form of supervenience, a logical or metaphysical supervenience, to be an essential aspect of what it means to be an emergent property.

High-level properties as aggregates

Kim also seems to want to preserve a sense in which higher physical levels have causal efficacy. He speculates that if causality can only happen on the rock-bottom level of microphysics, then one can perhaps always find smaller and smaller levels until "*-causal powers would drain away into a bottomless pit and there wouldn't be any causation anywhere*!" (Kim, 1998, p. 81). Kim thinks there is a tendency among anti-reductionist philosophers to interpret the physical domain excessively narrowly. I agree that a narrow interpretation of the physical domain is unnecessary.

Kim partly blames the standard micro-macro hierarchical model for encouraging the idea that the causally closed physical domain only includes the basic particles and their relations and properties. "*But this is a groundless assumption. Plainly the physical domain must also include aggregates of basic particles, aggregates of these aggregates and so on without end; atom, molecules, cells, tables, planets, computers, biological organisms, and all the rest must be, without question, part of the physical domain*" (Kim, 1998, p. 113). Here we see that Kim includes aggregates of basic particles into the physical domain, however it looks like Kim is committed to the idea that even though these aggregates can have relatively independent properties, these properties ultimately inherit all their causal power from the basic particles and strongly supervene on them.

The question then is if this interpretation is again too narrow to secure the higher level causality we want in order to secure free will. Kim would of course counter that the strong emergence I favor, would be a notion of physicalism that is overextended. In the end, physics will determine what is physical, and as we have seen in the past, new forces and fields have been introduced in ways that have been continuously surprising, upsetting and disruptive to the physics of the day.

Can aggregates count as causal properties?

In any case, the question for now is if Kim's proposal might provide the kind of high level causality I want here. Kim uses the following example to illustrate how an aggregate can have properties as a whole, that the individual base properties does not: "*Having a mass of ten kilograms is a property of certain aggregates of molecules, like my coffee table. And it is a micro-based property of the table in the following sense: for my table to have this property is for it to consist of two parts, its top and its pedestal, such that the first has a mass of six kilograms and the second a mass of four kilograms*" (Kim, 1998, p. 83-84). This table has a mass of ten kilograms and this property cannot be found in any of its parts, only in the whole table put together, in the same way, human beings have causal powers that none of our individual organs have (Kim, 1998, p. 85).

According to Kim, just because the causal power of the macro-property can be explained by its micro-structure, that does not mean that its causal power is identical with the causal power of its micro-structures. They are new causal powers that add to the causal structure of the world. This is not in conflict with the basic commitments of physicalism and physicalism should not be identified with micro-physicalism (Kim, 1998, p. 117).

Kim goes on to conclude, on the basis of these examples that "*Clearly then, macro properties can, and in general do, have their own causal powers, powers that go beyond the casual powers of their micro-constituents. This is an obvious and important point to keep in mind*" (Kim, 1998, p. 85). Now, in what *sense* are these genuine higher-level causal powers?

One problem with such a weak account seems to be that if aggregate objects have macro-causality just as a result of the joint micro-causality of their aggregated base parts, then there is no special kind of causation that happens at the higher level in particular instances. If the coffee table has causal power, then so do all conceivable assemblies of physical aggregates, with no aggregates having any special status above any other.

For Kim then the causality at the higher level is purely in virtue of the physical law-like regularities happening on the fundamental physical level. But if we wonder how to explain the "in virtue" relation, how can we outline different philosophical positions? How do we specify in detail the clear account of strong emergence that we want here?

Logical, metaphysical or nomological Emergence?

One way to think of the "in virtue of" relation is to note that there are different senses of emergence, relative to the kind of relation that holds between the lower and higher levels. As noted above, this relation can be logical, metaphysical or nomological. In terms of a logical supervenience-relation, it means that we can deduce from the lower levels all possible higher levels, a priori, at least in theory. For nomological supervenience on the other hand, this "in virtue of" relation would not hold either logically or metaphysically, and so could not be known a priori, or purely philosophically. In other words, while five stones individually and the laws they are governed by are empirical and nomological, the subsequent deduction required to understand five stones as an aggregate in unity, is logical and mathematical, and requires no further nomological premises. The contrast between this weak sense of emergence then, and a strongly emergent view, is that further nomological premises must be added to bridge the lower and higher levels in the strong emergence case. In that case, some combinations would be special, as compared to mere aggregates, in requiring extra nomological laws.

To give an example, Chalmers thinks that Zombies are logically *and* metaphysically possible, but not nomologically possible (Chalmers, 1999 in Kirk, 2019). If this is true, it means that there cannot be a logical supervenience relation between the mental and the physical of the kind Kim assumes, because the same lower level physical state can give a different mental outcome in a different world, and so cannot be a logical or metaphysical supervenience. This is not an argument against Kim, but to show that a stronger sense of emergence is needed to explain the sense of free will I'm looking for.

Emergence and holism

Another problem with Kim's view, insofar as we wish to account for consciousness and free will, is that it does not include holism. Holism is the idea that the whole is greater than the sum of its parts. This seems to be an important aspect of emergent phenomena like consciousness. As Humphreys puts it, "*The idea is that, in contrast to emergent entities, aggregate entities are nothing but their constituent units, a particularly clear example being the building material called aggregate, which is simply a collection of small stones used in*

making concrete. Aggregate entities can be structured, but this structure is not permitted to lead to the development of a distinctively new kind of property." (Humphreys, 2006, p.37). In other words, the problem with many of Kim's aggregate examples is that the structure of the whole does not lead to the development of a distinctly new kind of property existing over and beyond its parts.

Kim would argue that his view of aggregates extends to cells and other entities that do seem to lead to distinctly new kinds of properties. However, it is not clear how these examples of distinct new kinds of properties follow from considerations that apply in the simple cases, with stones, tables and other aggregates. In addition, since Kim's definition of emergence is logical and metaphysical, rather than nomological, such new properties should be derivable in theory, even if they may be unexpected in practise. So unless Kim can demonstrate that such distinctive new kinds of properties follow logically and metaphysically from nomological premises that are already established, or the strong emergentist can demonstrate that they do not follow logically and metaphysically, then it seems we will be at an impasse.

The main worry for reductionist physicalists like Kim seems to be that if we allow for the existence of strongly emergent phenomena that cannot decompose into fundamental physics, this automatically means that we have to abandon physicalism and succumb to dualism; "*Leaving physicalism behind is to abandon ontological physicalism, the view that bits of matter and their aggregates in spacetime exhaust the contents of the world"* (Kim 2005, p. 71). However, as a strong emergentist with a physicalist bent, I do not believe that this is a good or final definition of physicalism as a scientifically inspired philosophical view.

Further, Kim supposes that a strong emergentism would entail *"embracing an ontology that posits entities other than material substances-that is, immaterial minds, or souls, outside physical space, with immaterial, nonphysical properties*" (Kim 2005, p. 71). Again, this idea that anything but classical physics plus aggregates would entail what is essentially magic, does not hold up under scrutiny, and is in danger of projecting a failure of imagination into the fabric of the universe. Historically, science has repeatedly expanded to include things like electromagnetism, warped spacetime, and quantum measurement.

This is not to say Kim is wrong, but to preclude the possibility of strong emergence in such a dogmatic way should require extraordinary evidence, demonstration or philosophical proof. While Kim to some extent attempts to do this, I find his arguments inconclusive so far, and so continue on my quest to seek a nomological supervenience that can serve as foundation for

free will. Ultimately, I do not seek here to argue for the impossibility of Kim's view, but rather to argue for the possibility of my view, to carve out the possibility for robust free will.

Next we will look closer into the idea of strong emergence and see what problems we face by embracing this view.

Strong emergence

Strong emergence is roughly the idea that higher-level properties are only weakly dependent on, or determined by, lower-level properties. This means that the lower-level properties only determine the higher-level properties nomologically and not logically or metaphysically.

Strong emergence rejects strong reduction of emergent phenomena and accepts fundamental higher-level causal powers. Strong emergence is a general claim, but is often associated with arguments that the conscious mind must be strongly emergent in relation to its neural substrate (O'Connor, 2020). To clarify the notion of strong emergence, I will explain Chalmer's view on it. Chalmers thinks that strongly emergent phenomena exist and that consciousness is a clear example.

A system is typically considered conscious when there is "something it is like to be" that system (Chalmers 2006, p. 3). Since we are such systems, it is apparently a fact that they exist, although it is disputable whether or not they are strongly emergent. Chalmers also thinks we have good reason to believe that consciousness is not reducible to physical states. Nevertheless, since conscious states are correlated with physical states, it is plausible to assume that the brain is responsible for our conscious states (Chalmers 2006, p. 4). In this world it looks like identical physical states will create more or less identical mental states. However in other worlds with the same low-level physical laws, there might be physical systems identical to our brains that have no consciousness at all (zombies). This means that the lawful connection between physical states and mental states might not be derived from the laws of physics alone but need to include its own basic laws or some other basic law. Chalmers suggests we might call such "vertical" nomological laws for "fundamental psychophysical laws" (Chalmers 2006, p. 4).

It would take us too far afield here to go deeper into the meaning of Chalmer's notions or the notions of metaphysical, possibility and physical, and conceptual possibility. The main point for now is that Chalmers thinks of strongly emergent phenomena as "*naturally but not logically supervenient on low level facts*" (Chalmers 2006, p. 4). In this sense of strong emergence, Chalmers thinks fundamental physical laws must be upgraded with other fundamental laws in order to explain strongly emergent phenomena like consciousness.

There are several problems with strong emergence that must be solved for it to be a tenable view. I explain the main problems I see in the next section, and then in the next chapter I will go into depth by giving an account of a kind of strong emergence that solves these problems, namely fusion emergence.

Problems with strong emergence

There are quite a few problems with strong emergence that we can present by noting the worry that arises with regards to supervenience and causal closure. To clarify the situation, we can list the worries in the following way:

1. **Supervenience Worry:** If a strongly emergent phenomena at time t is strongly distinct from the physical, in the sense of not being metaphysically determined by the occurrence of a physical base feature at that time, we might worry that this definition of the strongly emergent misses something essential to the very notion of emergence.
2. **Causal Closure Worry:** Strongly emergent phenomena, in being strongly efficacious (in addition to strongly distinct), seem to break the causal closure of the micro-physical. If the strongly emergent phenomena are efficacious then there seems some sort of downward causation must be possible, which seems strange.

First, the supervenience worry is the worry that a mere nomological necessity is not sufficient to secure an emergent phenomena. For if it is *merely* a nomological necessity, then there is always the metaphysical possibility that e.g. zombies are realized in other worlds. This would mean that there is no conceptual or logical way to reduce a phenomenal property to the base layer of physical properties. That entails that the emergence has to be a brute fact of this world in particular, which seems to be an unsatisfying explanation. However, if we compare it with nomological laws in general, it fits into the same frame we use when we introduce

new fundamental laws to explain a wide variety of physical phenomena, but are themselves not necessarily given an explanation.

Second, the causal closure worry is only a problem so long as the strongly distinct emergent property is denied a fundamental physical law like a "fundamental psychophysical law" that will include bridge laws between itself and other fundamental physical laws. In this sense, causal closure is not broken, it is just extended to include higher level phenomena that have not yet been recognized by science as requiring new fundamental laws.

Note that the idea here is to show that this is a possible way to make room for free will, not to give prescriptions on how the physicists or psychologists are to do their job. If new nomological laws are required from the lower-level to the higher-level, it follows that it cannot be a philosophical job to find those laws, since it is ultimately an empirical question. Nevertheless, by fleshing out the conceptual possibility that there might be fundamental nomological laws that are "vertical" or "psychophysical", we don't rule it out a priori.

Especially when science is to concern itself with the seemingly problematic nature of free will, phenomenal properties, fusion causation and downward causation, it is good to keep possibilities open as to how they might be included in a broadly physicalist framework. To end this chapter I give a brief discussion of downward causation, before I proceed to lay down the positive view on these issues in Chapter 4: Fusion Emergence.

Downward causation

Downward causation means that emergent phenomena, in addition to being distinct and irreducible, also have causal efficacy. Happenings on the macrolevel can have causal consequences for other macro-level happenings or happenings on lower levels. Chalmers distinguishes between weak and strong downward causation, in the same way as with weak and strong emergence. With strong downward causation, high-level phenomena can have irreducible impact on low-level processes. With weak downward causation, in contrast, the causal impact of high-level phenomena are in principle reducible, but still unexpected (Chalmers 2006, p. 6).

The problem with strongly emergent downward causation is that it introduces fundamentally new arrows of causation which may create situations of overdetermination. These arguments are often referred to by appealing to “causal closure of the physical” or “exclusion”. While weak emergence can claim that the in principle reducibility of the high-level properties defends against these arguments, it is more difficult for strong emergence.

Thus it can seem that these arguments rule out all scientifically grounded possibilities for strong downward causation. In particular, strong emergence has often been brushed away as something mysterious and unscientific (Aharonov et. al. 2018, p. 11730). However, recently the eminent quantum physicist Yakir Aharanov, has claimed that there might be hope for strongly emergent properties that exhibit top-down causation. His research is based on the cutting edge of quantum physics, using his pioneering *two-state vector formalism* (Aharonov and Vaidman, 2008) and *weak values* framework (Aharonov et al. 1988) to discover novel phenomena in the heart of standard quantum mechanics.

In the next chapter we explore the philosophical position of fusion emergence, due to Humphreys, and the associated research of Aharonov, which attempts to ground fusion emergence scientifically by demonstrating an example of quantum mechanical fusion.

Chapter 4: Fusion emergence

Fusion emergence keeps the premises of reality, efficacy and distinctness, while modifying the premises of supervenient dependence and physical causal closure, in order to allow for strongly emergent phenomena. As we saw in the last chapter, strong emergence naturally leads to several problems that need to be addressed. Fusion emergence is a strong emergence view that aims to solve these problems in a positive way. Fusion emergence was first advanced by philosopher Paul Humphreys in his paper "Aspects of Emergence" from 1996. His latest book on emergence takes up the idea of fusion emergence in more detail in 2016, and more recently, Aharanov in 2018 attempts to scientifically ground Humphreys philosophical position of fusion emergence by appealing to quantum physical phenomena.

Fusion emergence

Fusion Emergence is a philosophical position aimed at solving some of the problems of strong emergence, particularly overdetermination in cases where strong emergence implies a kind of downward causation, from the whole to parts. Fusion emergence has also been used in a scientific context to make sense of how particles that exist on lower levels fuse together in forming new emergent properties at a higher level of complexity, where the lower-level particles that fuse together in a certain sense cease to exist (Aharonov et al. 2018, p.11730).

While we cannot fully go into detail on Aharanov's proposal, as it presupposes an in-depth understanding of quantum mechanics and information theory, we will explain the proposal in brief to give sufficient background for philosophical reflection. The idea here is that just as reductionism is an inherently philosophical position that often appeals to scientific method and practise, fusion emergence is a philosophical position of strong emergence, that is similarly able to appeal to scientific method and practise at the most fundamental level. To distinguish the philosophical and scientific aspects of fusion emergence, we refer to Humphrey's Fusion Emergence (philosophical) and Aharonov's Fusion Emergence (scientific). First we will go into Humphreys account of fusion emergence and then discuss how Aharonov makes scientific use of this view.

Humphreys Fusion Emergence

One initial motivation with Humphreys fusion emergence was to show how supervenience relations fail to obtain when fusion occurs. The basic idea is that lower-level properties disappear when they fuse together, in order to produce a new unified whole, and thus over-determination is avoided (Humphreys, 2016, p.78).

Fusion emergence is a kind of transformational emergence. Transformational emergence suggests an ontological framework in which basic and structured properties undergo fundamental change, and thereby get new powers, that are not "latent" in the antecedent, ontologically-grounded base, and lose others. With the new powers, there are new laws describing their evolution. There is a constantly evolving dynamic, where elements are transformed through interactions with other elements (O'connor, 2020).

Humphreys (2016) describes transformational emergence in the following way:

> *Transformational emergence occurs when an individual a that is considered to be a fundamental element of a domain D transforms into a different kind of individual a*... as a result of interactions with other elements of D…They possess at least one novel property and are subject to different laws*…. (Humphreys, 2016, p. 60 in O'Connor 2020).

He takes an example from the Standard Model of particle physics, which describes partless muons as very quickly "decaying" into electrons, electron neutrinos, and muon neutrinos (Humphreys, 2016, p 66–67 in O'Connor 2020). Here there is fundamental change, not merely a change within individuals but change *of* individuals from one kind to others. Humphreys suggests that fusion emergence is a special case of transformational emergence. When fusion occurs, basal entities or certain of their properties are lost when they fuse with others in producing a unified whole (Humphreys, 2016: 74–5 in O'Connor 2020).

The basic idea of Fusion Emergence according to Humphreys "*is that two property instances that belong to a domain D interact, and in so doing, the instances are transformed in such a way as to produce a new property instance, the key feature of which is that it does not have the original property instances as components"* (Humphreys, 2016, p.75).

Humphreys tends to use examples drawn from fundamental physics, concerning for instance the transformations of the muon in elementary particle physics, or "*the phase transitions that give rise to superconductivity and superfluidity in helium are a direct result of quantum entanglements*" (Humphreys, 1996, p.66), as concrete physical instances of how fusion emergence may play out in the scientific, empirical realm as actual fact (Humphreys, 2016, p. 66). In this way, he also anticipates Aharanov, in already talking about quantum entanglements as problematic to supervenience.

Humphreys fusion example

Humphreys gives the general example that "*one of the distinctive features of quantum states is the inclusion of nonseparable states for compound systems*" (Humphreys, 1996, p. 66) such that we have the following situation:

Nonseparable compound systems: "*Within these states, the composite system can be in a pure state while the component systems are not and the state of one component cannot be completely specified without reference to the state of the other component. Moreover, the state of the compound system determines the states of the constituents, but not vice versa.*" (Humphreys, 1996, p. 66)

A pure state in quantum mechanics is described by a vector in Hilbert space. It is the mathematical operations on such a vector, and the eigenvalues of those operators, that is associated with measurement and the possibilities of measurement outcomes respectively. However, a technical discussion will take us too far afield here, the main point is to compare in relative, conceptual terms this situation with the definition of supervenience used above:

Basic supervenience: ***"****If property M emerges from properties N\,...,Nn, then M supervenes on N\,...,Nn. That is to say, systems that are alike in respect of basal conditions, N\,...,Nn must be alike in respect of their emergent properties*." *(Kim, 2006 p. 550).*

The problem with contrasting nonseparable compound systems and basic supervenience is clear, even without a deeper technical understanding. The very basal conditions that are

presupposed unproblematically in basic supervenience, cannot be known individually in nonseparable compound systems apart from their determination through the whole system. In a recent article by Mičuda et al. (2017) they put it like this: when ".. *well defined and distinguishable local properties are superimposed*" into nonseparable compound systems, then ".. *the individual properties are smeared out whereas the state as a whole still exhibits well defined global properties*" (Mičuda et al. 2017, p. 1). Thus, there seems to be some loss or "smearing" in the low-level properties while there is an irreducible holism of the whole.

It seems then that the advent of the emergent higher level holistic property comes at the cost of the lower level properties. Quantum physicists are apt to explain this "cost" epistemically, in terms of knowability, or information, in the sense that because we know more about the whole, and how that in turn determines a restriction on what the parts can possibly be, we know less about what the parts are individually. Considering the close ties between epistemology and ontology in quantum mechanics, it is not a simple matter to question the ontological implications of such a result on those grounds alone, though that is certainly a possibility. To understand causality at the quantum mechanical level, is to understand what informational transformations are possible and which are not. Quantum information theory, for instance, helps specify the range of possible operations of a quantum computer, and so also specifies the possible physical transformations that are possible. However, it would take us too far afield to attempt a justification for the inevitable link between epistemology and ontology in quantum mechanics.

Recalling Humphreys notion of fusion emergence then, when some local property instances of microphysics interact, and in so doing give rise to a new property instance, the new property instance has strongly emerged by fusion into that new property instance. When the new property instance has emerged, the microphysical properties have become "smeared out" and lose some of their properties. Exactly in what sense they are smeared out will differ depending on the instance of fusion, but in the quantum mechanical sense it seems they lose some of their informational properties, which is closely linked to their ontological status. In a sense, we can say that some low-level properties in fusion can be traded for a higher-level property in a way that is nomologically, but not logically determined.

Now, given that such a view can be defended, how does it help solve some of the problems we have discussed? The idea is that fusion emergence will help us formulate a view of strong emergence that can satisfy our supervenience and causal closure worry described earlier.

How fusion emergence solves the supervenience and causal closure worries

Supervenience Worry

The supervenience worry takes on a new form in light of fusion emergence. Recall how we formulated the supervenience worry earlier: If a strongly emergent phenomena at time t is strongly distinct, in the sense of not being metaphysically determined by the occurrence of a base feature at that time, we might worry that the emergent relation misses something essential to the very notion of emergence. What to make of this in light of fusion emergence?

The fused property emerges because it results from an interaction between the original properties: "*A holistic element is present in the fact that the fused properties form a whole without identifiable components"* (Humphreys, 2016, p. 78). Because of this we need to take care to distinguish, as does Humphreys, between synchronic and diachronic emergence, where the former typically means the coexistence of lower-level components at a particular time t, while the latter is more general, with the emergent property emerging from some process (Humphreys, 2016, p. 67). We thus get two senses in which the emergent property is said to depend on original properties.

In fusion emergence, since the original properties that interact lose some of their essential properties in the process of fusion, it cannot be true that they are synchronically coexistent with the emergent property (because by that point in time they have already lost some of their essential properties), but it can still be true that the original properties diachronically gave rise to that fused property. The understanding that supervenience can hold diachronically as well as synchronically is the main way to help alleviate the supervenience worry.

In addition, there seem to be no special problems with taking emergence to be nomological as opposed to just logical or metaphysical. In terms of giving a good explanation of a higher

level phenomena, one may feel that the explanation would be better if it was completely reducible to the lowest possible level, however, from a scientific point of view, one has to postulate laws somewhere, and a "psychological" law is in principle no different from a law in fundamental physics, as giving nomological conditions on what follows from what. The requirement that it is hard to reject, however, is that such laws interact with and be consistent with other natural laws (causal closure) but that is exactly what fusion emergence accepts.

Causal Closure Worry

The causal closure worry also takes on a new form in light of fusion emergence. Recall the formulation of the causal closure worry earlier: Strongly emergent phenomena, in being strongly efficacious (in addition to strongly distinct), seem to break the causal closure of the micro-physical. If the strongly emergent phenomena are efficacious, then there seems some sort of downward causation must be possible, which appears strange.

The key solution and problem here with fusion emergence is that *"the disappearance of the original property instances renders the fused property autonomous from those original properties*" (Humphreys, 2016, p.78). In other words, by supposing that a completely distinct property instance arises, there is a new domain of causality that goes beyond what is synchronically available in the base after fusion. What is synchronically available in the base after fusion are those properties that have been "smeared out" or lost some of their properties.

The fact that a fundamental entity can be transformed, makes fusion emergence essentially different from generative atomism, in which the fundamental entities are fixed (Humphreys, 2016, p.60). *"When the new instance belongs to a different domain D' and the exclusion problem has to be addressed, the problem disappears because there is no overdetermination between the causal processes in the original domain and those in the new domain.*" (Humphreys, 2016, p.75). From the point of the old domain then, there has arisen a new domain that challenges the causal closure of the old domain, because it contains causal powers that are not synchronically reducible to the old domain. However, if we extend our scientific domain to include the new domain, by some new "vertical" nomological laws, there is no problem with causal closure from a scientific perspective.

The (scientific) difficulty now instead lies in understanding how the emergent properties of the new domain interact with the old domain, and that is mainly a scientific question, as it occurs in each instance. We will not go into this here, as it is mainly a scientific question, but let me just give an example to make it clear how this might go.

Think of the case of nonseparable compound systems as discussed above. It seems that in order to determine certain states that the parts are in, one must go through the whole system. And so the causal power one gets over the parts by going through the whole, seems to be different than the causal power one can get by going through each part individually. At the same time, there is no overdetermination, because there is a precise balance, in that what one has gained by being able to determine the parts through the whole, one has lost by not being able to go through each part individually. Quantum mechanically, according to Aharanov, it seems that one can arrange the system in a certain holistic way to get information about each part, that is impossible even in principle to get by arranging each part separately to give out the information. At least that seems to be the gist of it conceptually. To go through the quantum mechanical mathematics of this would again take us too far afield, and the technical details would depend precisely on a given instance of fusion emergence in nature.

This conceptual gist of such a quantum fusion view may also serve as an analogy to the qualia fusion view that we discuss below; the idea in the qualia case being that we as epistemic subjects get access to certain kinds of information through the whole of experience, that a third party cannot get by simply deriving it from the brain, part by part. In either case, it seems that these would be subtle effects that are not easily discoverable, but would require not only conceptual but technical and practical solutions that only hard science can provide.

Before we get into the conceptual problems with Fusion Emergence, let us sum up some of the important conditions and entailments of fusion emergence that Humphrey lists.

Conditions and Entailments of Fusion Emergence

Humphrey outlines four conditions of fusion emergence, that it is; relational, novel, autonomous and holistic. The condition of being relational means that emergent entities must result from something else (Humphreys, 2016, p.28). Novelty, means that an entity is impossible to deduce from a theoretical base using the theoretical apparatus of that base (Humphreys, 2016, p.28). The autonomy condition is that the emergent property exists independently of its base (Humphreys, 2016, p.33). Holistic means that emergent features are "more than the sum of their parts", they are not aggregates (Humphreys, 2016, p.36).

Humphreys also mentions that fusion emergence entails that; *"There is a dynamic or quasi-dynamic aspect to the fusion process in the sense that fusion results from a process of interaction and it is rarely instantaneous."* and that *"The result of fusion is an irreducible, unified, holistic entity."* and that *"Emergence can take place within the domain of the physical. This means that although the most fundamental domain is not causally closed, there need be no violations of physical law in virtue of that"* (Humphreys, 2016, p.78-79). Again the idea here being that the supplementation of additional nomological laws, as science has been doing for several hundred years, would not be the end of physicalism even if they appear now strange to what we are used to. So long as the new version is internally consistent and takes into account all phenomena, it would merely be another paradigm shift.

Next we will look at some problems that arise with fusion emergence.

Problems with fusion emergence

Several problems with fusion emergence have been detected.

1. The base properties lose some of their essential properties during fusion (this is called "basal loss"), such that these essential properties cease to exist. It is not clear how to understand the supervenience of the higher level properties when the lower level has basal loss has occurred.

2. Fusion Emergence implies higher-level to higher-level and downward causation which is not yet well understood. There needs to be a consistent set of rules for the interactions during and after fusion emergence between higher and lower levels.
3. Humphreys fusion emergence proposal may seem unscientific or mysterious, as it hasn't been recognized by mainstream science yet, in contrast to most reductive atomistic views.

In regards to the first problem, Humphreys replies that one could distribute the parts in a delicate balance so that some properties at the lower level could fuse and some could keep their identity as individual parts to maintain the structural integrity of the whole. The idea being that while some of the properties of the base cease to exist, in giving rise to the fusion emergent properties, other properties keep their existence to support the fusion emergent property as a base.

It is not completely clear what Humphreys means here, but one may imagine a metaphor wherein one uses *some* of the material of some object to create a novel extension of that object, while leaving *some* materials for support in order to keep it steady and upright against gravity. In such a case, the metaphor is that some of the material will be transformed into the new fusion-emergent property, while some of the material (lower-level properties) will remain as they are for support. In either case, Humphrey has gotten some resistance on this, for while in theory it may be made to work, it seems conceptually difficult to do so, and the view may seem poorly motivated.

Manafu, for instance, holds that Humphreys' distribution of labor between properties would not solve the problem of basal loss if we do not have *independent justification* for why the division between properties that can undergo fusion and those that cannot, should overlap with the division between the properties that are not essential in the functioning of the system and those that are (Aharonov et al. 2018, p. 11731). Second, in addition to lack of independent justification, fusion emergence is a matter of *such delicate balance*, that it is difficult to see how to balance the "loss" of the basal properties, with the "gain" of the emergent property.

In regard to the second problem, there is simply a lot of work to be done, not just with regard to downward causation, but also with high-level to high-level causation. In some sense one has to solve both problems at the same time to plausibly give an account of how the emergent properties have causal power without violating the structural integrity of the system under analysis, or fundamental law of physics. That a position is conceptually difficult, however, should not detract from the question of its truth.

In regards to the third problem, the best solution is to present a scientific project in fundamental physics inspired by and supporting fusion emergence. This will also independently motivate the view, and provide the required technical ability to balance the basal loss with the emergent gain, without violating any fundamental physical laws.

Aharonov et al. appears to provide exactly such an independent motivation to pursue fusion emergence from the perspective of fundamental physics, while also solving problems with fusion causation by giving us a concrete example of fusion from quantum physics.

Aharonov, famous for the Aharonov–Bhom effect, (Aharonov and Bohm 1959, p. 485-491) and its dual the Aharonov–Casher effect (Aharonov and Casher 1984, p. 319-321) appropriately represents a paradigm shift within quantum mechanics, as he has helped pioneer the "two-state vector formalism", an interpretation of standard quantum mechanics in which the present is understood as being caused by quantum states of the past and of the future taken together (Aharonov and Vaidman, 2008, p. 399-4447). With this framework Aharanov and his colleagues have discovered a range of surprising effects and phenomena; Weak Values (Aharonov et al. 1988), Quantum Cheshire Cats (Aharonov et al. 2013), Superoscillations (Aharonov et al. 2017), and Negative Kinetic Energy (Aharonov et al. 1993). The phenomena that concerns us here, is his article on the "Completely top–down hierarchical structure in quantum mechanics" (Aharonov, Cohen, and Tollaksen, 2018). Given Aharonov's impressive track record in the field of fundamental physics, it is worth taking into consideration his engagement with Humphreys philosophical views, and his views on how fusion emergence can be realized in an actual physical instance.

Fusion Emergence in Quantum Physics

As we saw above, Humphreys notes that the non-separable states of quantum mechanics give rise to examples of fusion emergence. These anticipatory remarks by Humphreys have been followed up by a recent paper by the eminent quantum physicist Aharanov. In "Completely top–down hierarchical structures in quantum mechanics" (Aharonov et al. 2018, p.11730) Aharonov outlines Humphreys view and creates his own summary which states that Fusion Emergence has at least three important consequences (Aharonov et al. 2018, p.11730):

1. "*The whole can not be reduced to the parts (the parts no longer exist after fusion).*
2. *The top–down causal efficacy of the "whole" is not in conflict with the causal efficacy of the original lower-level properties which radiate their causal efficacy in the usual bottom–up fashion.*
3. *The new emergent properties can now have causal efficacy over the parts*" (Aharonov et al. 2018, p.11730).

Aharonov et al. then proceeds to show how these three consequences are satisfied in experimental conditions, by appealing to the following quantum mechanical experiment:

> *"Consider a conceptual example with three particles and three boxes. With appropriate preselection, postselection, and measurement during the time after the preselection and before the postselection (below), we observe the following properties: No individual particle is found in any individual box, no two particles are found in any two of the boxes, and no correlations are found between any two of the boxes. However, when we measure the correlation between all three boxes, we nevertheless find surprising, strong correlations. We could not use the one particle or two-particle information to deduce the properties of the three-particle correlation. However, with the information of the three-particle correlation, we can now, in a top–down fashion, deduce the one- and two-particle properties and correlations."* - (Aharonov et al. 2018, p. 11731).

According to Aharonov, this is a general phenomena, consistent with the known, standard laws of quantum physics. That is, according to Aharonov, we can find higher-level correlations that cannot be derived from the lower-level correlations, whereas the lower-level correlations can be derived from the higher-level correlations (Aharonov et al. 2018, p. 11731). While Aharanov frames the result in terms of information, as noted, this is how scientific results with clear ontological import are often framed in quantum physics.

In theory fusion emergence might be an elegant solution to the problem of mental causation. It respects both the existing laws of nature that have given rise to our intuitions about physical determinism, while also opening up the possibility that the mental has causal power through fusion emergence. According to Humphreys, current evidence suggests that fusion

emergence is not uncommon in the physical and chemical domains, but argues that it certainly does not happen everywhere. Similarly, Aharanov's argues that the experiment can be generalized, but that it nevertheless requires a delicate set-up.

Problems with Quantum Fusion Emergence

Several problems with quantum fusion emergence have been detected.

1) It is not clear that Humphreys novelty premise is satisfied.
2) It is not clear that we need any new laws to explain this kind of emergence.
3) It is not clear that this emergence is ontological, rather than epistemological.
4) It is not clear that it can help resolve the problem of strongly emergent mental states.

Regarding the first problem, Humphreys novelty premisestates that an entity is impossible to deduce from a theoretical base using the theoretical apparatus of that base (Humphreys, 2016, p.28). However, Aharonov seems to be using the theoretical foundation of quantum mechanics to simply deduce the effect. Therefore, it seems quantum fusion causation is not a process satisfying the novelty premise, and thus not fusion emergence as Humphreys defines it. There are two ways to approach this objection.

The first option is to reply that the novelty premise is not broken, and so save Humphreys definition of fuson emergence. One can say that although quantum mechanics predicts irreducible effects, this does not mean that it can simply deduce what those effects are. In simple cases this may be so, but in more complex cases, the irreducibility is also an irreducibility of our ability to deduce the effects. Ultimately, if we need to instantiate a system of the same complexity (for example a Fusion Quantum Computer) as the system we wish to understand, it is not clear that we have reduced it, but merely emulate it.

The second option is to reply that Humphreys novelty premise is too restrictive. It is not always clear what amounts to a "theoretical apparatus", in the sense that if the apparatus is complex enough, for example some sci-fi Fusion Quantum Computer, the distinction between the "theoretical" apparatus and the entity we study begins to dissolve. If a sufficiently capable

Fusion Quantum Computer can actually become conscious for instance, we don't necessarily have to think that it was reducible after all, as little as my having a certain pain is a reduction of the pain of someone else.

That the fusion emergent entity lies as a "seed" in the substrate and properties that undergo fusion, is not enough to count against strong emergence, so it makes sense to say the same with regards to the theoretical base and apparatus. In the end the theoretical base and apparatus have to reflect that base in a true way. The main problem is rather the second objection 2), that it seems we need no *new* laws to explain quantum fusion emergence. Recall that Chalmers held a view of strong emergence that only entails a nomological necessity from the base to the emergent properties. This means that if we have captured all the laws in quantum mechanics, the emergent properties follow with nomological necessity (although not metaphysically or logically).

However, we only suspected that the emergence of the mental requires new laws, but if the mental emergence is just some high level aspect of quantum fusion emergence, then that is the way it is. In either case, we are committed to broadly physical laws that entail mental fusion emergent properties nomologically. This is why we cannot accept a version of the novelty premise that does not include the possibility of the emergent entity being there as a seed in a final theory. In either case, this kind of fusion emergence would be much stronger than accounts of emergence where the supervenience entails a fixed metaphysical dependence from lower to higher, as some of the lower-levels properties cease to exist in fusion.

Another worry is objection 3) that the quantum mechanical fusion result is epistemic rather than ontological. This is too intricate a question to go into in this thesis, as I have mentioned, because quantum physics in general presupposes an intimate link between the epistemological and ontological situation, for example through the uncertainty principle and various information-theoretic results.

Finally, in objection 4) I ask, how can quantum fusion emergence help explain the strongly emergent mental states? It seems unlikely that the mental is a result of quantum mechanical effects like the one used as an example for quantum fusion emergence. Humphreys himself holds that the mental domain is currently not the best place to look for examples of

emergence. According to Humphreys, neuroscience does not suggest the existence of autonomous, fusion emergent mental states and properties (Humphreys, 2016, p. 89). Despite showing that fusion emergence is consistent and plausible, it is another matter now to extend that view to help account for the causal power of mental states. Theories to the effect that the mental requires strongly emergent properties for mentally relevant effects, are still speculative, and the issue is still controversial. However, it is too early to conclude that the current evidence excludes the possibility that the mental makes use of fusion emergence to instantiate strongly emergent causal properties. In any case, the possibility remains open so long as the view is consistent, and no evidence currently contradicts it.

In order to find a satisfactory solution that avoids overdetermination and dualism of the mental, we find that the best option is to explore mental fusion emergence. Fusion emergence as we use it is committed to the causal closure of the physical, and so extends the domain of the physical to preserve this principle. In this way it avoids overdetermination and dualism of the mental/physical, at the cost of introducing a conceptually intricate notion of emergence as we have seen. To do that for the mental domain, however, we need to zoom out a little bit, to get a philosophical overview of the debates about strongly emergent mental states and how fusion causation might be relevant.

Fusion emergence and the mental

Why do we need fusion emergence to be able to make sense of the mental? Is it not enough to give a plausible account of how the mind can weakly emerge in the way suggested by Kim earlier? If neither Humphreys nor Aharonov seem to have any intention of applying fusion emergence to the mental, why do we think it is worthwhile to pursue this approach?

As we will see in this chapter, I think the mental is strongly emergent in a way that supplements my libertarian view of free will. The belief in free will, coupled with the rejection of determinism and epiphenomenalism seem to require strongly emergent mental properties that can account for free will. In addition, my rejection of overdetermination and by embracing the causal closure of the physical, I am left with fusion emergence in order to stay within the bounds of a broadly physicalist account of free will.

Of course this is not the most common of views. According to Kim, for instance, weakly emergent aggregates can inherit causal efficacy from the fundamental physical domain in the following way: *"According to the causal inheritance principle, the causal powers of an instance of a second order property are identical with (or a subset of) the causal powers of the first order realizer that is instantiated on that occasion"* (Kim, 1998, p. 116). This means a kind of weak emergence, where the emergent property has its own causal power. But this power does not go beyond the powers of the first-order properties which are the constituents of the emergent property. For Kim, there is no special problem with weakly emergent properties and if mental properties turn out to be weakly emergent in this way, there is no special problem about the causal role of the mental.

Is Kim's causal account enough to create mental causation? It depends on what kind of mental causation we want. The main problem with the aggregate account is that it does not solve the problem of overdetermination in the way we wanted by accepting irreducible high-level causes that come into existence only at the cost of a "basal loss" meaning that some lower-level properties objectively cease to exist. The aggregated macro property can only have causal power through its parts in a bottom-up fashion, it cannot give any plausible account of downward causation that is irreducible to its parts.

The way I see it, I can accept aggregates and most of the things Kim says, but also hold that there is a special kind of fusion emergence that can happen, in a delicate way in certain situations, that can not be reduced to aggregates. In contrast to aggregated causal power, with fusion emergence, some combinations effectively have an associated "cost" in that some of the essential properties of the base layers will cease to exist before the new can emerge. This supports the notion of "costly unity" with regard to the mental, because it means that the fusion of an entity implies the actual loss of some properties and the emergence of others. In contrast aggregating something in this way or that way does not fundamentally change the existence of the parts that are aggregated.

Therefore I suggest at least three important aspects of mental causation that Kim's account does not provide; costly unity, no systematic overdetermination and downward causation. Since we are going to end up with a view of consciousness in particular, rather than the mental in general, we talk here in terms of consciousness.

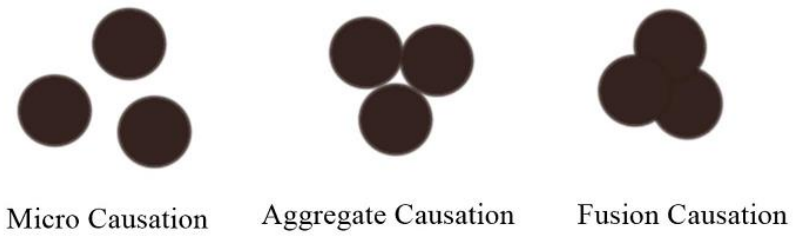
1. **Costly Unity.** Why must consciousness be costly unified? First, from a first person perspective it seems obvious that my consciousness is unified, it cannot be open to interpretation whether I see a unified field or not. In a particular moment it is the case that either I see the whole room, listen to music and taste coffee at the same time, or I do not. All these experiences can exist together, in the one single unified field of consciousness, in that particular moment in time. Second, it seems clear that my consciousness has a cost related to the unification; for instance, the field of consciousness has to relate all the things within my consciousness, from a particular perspective, with certain colors, things related a certain precise way, etc. As purely existent, A and B may be related metaphysically free of charge, but when they are related in a subjective perspective (and that subjective perspective is understood to be physical) it plausibly carries a cost. This is analogous to the quantum mechanical case, where the information at the higher level carried a cost that was paid by the lower levels being "smeared out." Generally we don't think of consciousness as having any cost, because we generally don't take seriously the idea that it is a genuinely physical phenomena, with genuine causal power. However, I notice that as soon as I seriously entertain the idea that consciousness is genuinely physical, strongly emergent, and causal in every spec and dust of its phenomenal properties, it is suddenly much more plausible to suppose that any property or aspect of consciousness has an associated cost that must broadly be understood in physical terms. For instance, any physical property must respect causal closure, and that means conservation of information, conservation of energy etc. Similarly, fusion emergent mental states, if they are broadly physical, must also respect these conservation laws or some not yet understood extension to such conservation laws.

2. **No Systematic Overdetermination.** As we have discussed earlier, in order to have strongly emergent irreducible mental causation, the mental cause cannot systematically compete with the underlying micro-physical cause for causal efficacy. It must be a genuine macro-cause in its own right. The aggregate causal account does not give us this possibility, if we argue for this account we must accept that any macro-property gets its causal power from a relatively straightforward combination of lower-level constituents, in order to avoid overdetermination. Fusion emergence can help us solve this problem, as we have seen, where it is not just that it is a passive

costly unity, that emerges just to give information to a subject, but something that at the same time has an irreducible, strongly emergent causal power that gives way to an irreducible kind of downward mental causation.

3. **Downward Causation.** Consciousness must have a kind of downward causation that is not a summary of the causal powers of its parts, but a power in its own right. It is widely claimed that the human experience of conscious choice is a case of direct, "top down" control over one's action by the agent (O'Connor, 2020). As we will see below, the phenomenal properties argument, and the evolutionary argument force us to strongly consider this possibility. These arguments use the idea that if we take phenomenal consciousness at face value as irreducibly existing at a high level, and reject epiphenomenalism, then we can argue for its causal power, which is plausibly understood as a kind of downward causation in light of a fusion emergence account. For now, it seems clear to say that we experience ourselves not only as the source of psychological and physical influences that converge and result in behavior, we also believe that we have the power to decide which options we shall take in everyday life.

Regarding downward causation, the picture below points out the main difference between normal micro-physical causation, the aggregate causation suggested by Kim, and fusion causation suggested by Humphreys.



Even though aggregated causation goes beyond the causal powers of its micro-constituents, it does so only as individual micro-constituents glued together. Combining micro-constituents in this loose way might not be enough to make them interact and integrate in the complex way needed to create consciousness. In particular, it does not seem to have an associated cost, as the micro-constituents keep all their properties intact. Aggregated micro-parts can only radiate their causal power in their usual bottom-up fashion and it seems that this is not enough to give a plausible account of mental causation. I think mental causation requires genuine top-down causation, where the mental cause acts completely independently of its

micro-physical parts, and that the micro-physical parts somehow pay the cost in fusion. And that's why I want to further explore the kind of causation we get with fusion emergence. But first, let us see exactly how fusion emergence can help us solve these three issues.

Fusion emergence provides costly unity, no overdetermination and downward causation

1. **How does fusion emergence provide costly unity?** As we saw from Humphreys description, the result of fusion is an irreducible, unified, holistic entity that eliminates its constituents after fusion. This result is an ontologically costly unity that is not an aggregate of its parts, but a new, fused property that exists independently of its parts. In this sense, consciousness must be a fusion emergent costly unity, such that some of the essential properties of the parts objectively cease to exist (presumably some essential properties of the matter involved in creating my consciousness) as the whole objectively emerges (the presence of the whole consciousness and the various relations that obtain). In this sense fusion emergence seems to respect a certain kind of conservation law, which we must expect from any broadly physicalist ontology.

2. **How does fusion emergence avoid overdetermination?** As we have already discussed, the causal power of fusion emergent properties do not compete with the causal properties of its constituents, because some of the essential properties of the constituents are eliminated during the fusion emergence of the emergent property. Therefore we get a genuine high-level causal power that acts independently of fundamental micro-physics. Once we have this higher-level power, moreover, as it can be made broadly consistent with fundamental physics, we expand the domain of physics to include this domain, to respect causal closure and ultimately establish proper conservation laws. So strong emergence of consciousness has a cost, in that there is a price to pay for the material that consciousness emerges from. This balance must be described technically and ultimately scientifically, to ensure that the loss or "smearing out" of properties in the lower level, balances out with the gain in properties at the higher level.

3. **How does fusion emergence provide downward causation?** This is a difficult and complex question that quantum physicists like Aharanov are trying to solve. This is not to say that the mental is quantum fusion emergent, but we can learn from the approach of Humphreys and Aharonov. Their current suggestion is that properties of the micro-constituents are eliminated during fusion, their causal power disappears and goes into making the new causal power of the emergent macro property. This new high-level causal power is strongly emergent, but somehow it is actually able to affect happenings on lower levels as well as other high-level happenings. In this sense, consciousness has phenomenal causal powers through its costly unity, and through whatever relations hold in virtue of the properties of the whole, insofar as it has strongly emerged above and beyond the parts.

What is the difference between physical fusion emergence like the example above from quantum mechanics, and mental fusion emergence? Most of the fusion emergence examples have appealed to relatively simple quantum mechanical systems, and there is no evidence that such fusion emergence can happen at the psychological level to create consciousness. Nevertheless, the possibilities are there, and since the reward is so great, in terms of resolving a number of deep philosophical problems that plague the discussions about the mental-physical divide, it is worth exploring fusion emergence as a pioneering approach to problems related to the emergence of the mental from the physical. This does not mean that mental fusion emergence has to be the same as the quantum fusion emergence, only that it follows the general pattern of fusion emergence as discussed by Humphreys and Aharanov.

In particular, the distinction between mental and physical properties may become scientifically explicable if we use a fusion emergence framework, in the sense that one can strongly emerge from the other. In this sense we can maintain a proper distinction between mental and physical properties while also subscribing to a broadly physicalist framework.

Property dualism and fusion emergence

According to property dualism, the mental and the physical properties can be vastly different, and distinct from each other, while their fundamental base is ultimately the same. I suggest that this base is fundamental physics, where micro-physical properties can fuse, and thereby give up their existence to new, higher level properties that can include mental properties.

This can sound very strange and seem hard to grasp as a possibility, however if we make the analogy to electric fields and magnetic fields ultimately being of same force, namely electromagnetism, we see that two different properties, while excluding each other in any given instance, can be transformed into each other following specific laws as we arrange things in a certain way, with mathematical precision. Similarly if we imagine the brain being in an unconscious state, we can imagine one "field" of unconsciousness, and when the brain becomes conscious, we can imagine that unconscious "field" losing some of its properties (paying the cost) and thereby transforming into the conscious field. While either "field" excludes the other at a particular time, nevertheless they are of the same fundamental kind (similar to the electric and magnetic properties being of the electromagnetic kind) and might be transformed into each other while obeying general scientific conservation principles. Of course, any analogy like this will be incomplete, and the description of the electric and magnetic field may be better described as mathematical duals, so the analogy is not perfect.

More generally we have to imagine two properties, the properties of the constituents before fusion emergence, and the fusion emergent properties as they appear after fusion. Then the idea is that the properties of the constituents are transformed into the novel properties at the higher level. Thus, both properties are of the same conserved medium that admits the transformation. The cost is that in fusion some properties of the constituents vanish and give rise to the emergent property, and in fission the emergent property vanishes and gives back the properties to the constituents. Ultimately however, they are just properties of the same underlying medium, that admits this kind of a transformation.

Exactly how the mental and the physical properties are connected in actual fact seems to be a challenge for science. All I want to point out here is that science still has no solution to this problem, and that fusion emergence can be a pioneering new framework for understanding it.

In the next chapter, Chapter 5, we return to the task of showing how free will can be possible, now that we are armed with the possibility of strong emergence, through fusion emergence. In particular, we will focus on an essential component required for free will in the sense we argue for it here, namely non-reductive conscious causation.

Chapter 5: Non-reductive conscious causation

In this chapter I will give an account of non-reductive conscious causation, which will support our account of free will. First I will give a brief explanation of what I mean with consciousness. In this thesis, the most important aspect of conscious causation will be phenomenal properties, also called qualia. I present the phenomenal properties argument, which aims to prove that phenomenal properties have causal power. Then I present Hedda Hassel Mørchs evolutionary argument for phenomenal powers. Lastly I discuss how fusion emergence can help make sense of the causality of phenomenal properties, opening up the possibility for explaining strongly emergent conscious causation within the physical domain.

Consciousness

Searle, (2010) points out that consciousness definitely exists but that it is hard to define. However, if we are talking about a commonsense definition and not a scientific definition in terms of the most basic neurobiological processes of consciousness, then he finds it rather easy to give such a commonsense definition (Searle, in Baumeister et al. 2010, p 122).

Searle defines consciousness as the events, feelings, or awareness that we typically experience when we wake up from a dreamless sleep and that continues until we go back into dreamless sleep or become unconscious in other ways. The dreams we have while we sleep are in this definition a kind of consciousness (Searle, in Baumeister et al. 2010, p 122). I think that in the case of consciousness it makes sense to start out with such a common sense definition, because consciousness is something we all experience directly.

However, we also need to start characterizing consciousness to get a grip on how to think about it. According to Searle, the most essential features of consciousness is 1) that it is qualitative, meaning there is always a qualitative feel, or qualia, to any conscious state, and 2) it is subjective, since it only exists as experienced by a human or animal or possibly other kinds of subjects, and 3) it is unified in the sense that all of my conscious experiences (presumably at any given point in time, not across all time), are part of a single conscious field (Searle, in Baumeister et al. 2010, p 122). I agree with Searle that qualia, subjectivity and unity are essential features of consciousness.

These properties are also relevant in the explanation of our free will, in that we have a (qualitative) sense of freedom, that is presented to us as subjects (subjectivity), and that we have a view of a range of alternatives at a given time with the power to select among them (unity). Due to limitations in scope, we will mostly focus here on the extent to which qualia, or phenomenal properties, can have causal powers. This is in service of arguing for the possibility of strongly emergent free will, with fusion emergence as the explanation of how.

In the Phenomenal Properties Argument below I will argue that from the existence of phenomenal properties, it must inevitably have causal influence. I will also use the Evolutionary Argument to show why this is a plausible conclusion all things considered. Finally, I will show, given the acceptance that phenomenal properties have strongly emergent causal powers, that fusion emergence might be a good explanation of how it can be consistent with physicalism broadly construed.

The existence of qualia

Perhaps the most mysterious part of being conscious is our experience of "qualia" or "raw, sensory, feels" - the sensory experience of seeing the color red or tasting wine for example. In philosophical terms, what we mean generally with the term qualia here are phenomenal properties, meaning properties that characterize what it is like, or how it feels, for a subject to be in conscious states. The *"what it is like"* sense of consciousness was made famous by Nagel (1974 in Van Gulick, 2018), who argued that a creature is only conscious if there is "*something it is like*", for that creature to be in a mental state, like pain for example.

Access consciousness vs. phenomenal consciousness

In taking qualia, or phenomenal properties, to be the most important aspect of phenomenal consciousness, we thus implicitly take the "what it is like" sense to be the most important part of consciousness. However, some argue that there are aspects of consciousness, such as "access consciousness", that do not entail a "phenomenal" aspect. Ned Block (1995), defined access consciousness as having to do with intra-mental relations. In this sense a mental state

is conscious if it is available to interact with other states, and by the access that the agent has to its content. If the information in that state is richly and flexibly available to its containing organism, then it counts as conscious whether or not it has any qualitative or phenomenal aspect to it (Block 1995 in Van Gulick, 2018).

Similarly, it is not clear that all aspects of "conscious" cognition are phenomenal. For example, there is a question of what cognitive phenomenology (phenomenology of thought) would amount to, and to what extent thought, and conscious thought, essentially relies on a phenomenal aspect. In this thesis, I will only argue for the freedom of consciousness insofar as it involves qualia or phenomenal properties, and that any other freedom that concerns access consciousness or other non-phenomenal types of consciousness, must be argued for separately on other grounds, or in some sense be derived from the phenomenal type.

The phenomenal properties argument

The phenomenal properties argument is a line of reasoning that assumes the existence of phenomenal properties and the eleatic principle as true and concludes with the causal power of phenomenal properties. The phenomenal properties argument can be set up in the following concise way:

A. Phenomenal properties exist undeniably.
B. The eleatic principle is true.
C. Therefore phenomenal properties have causal power.

With this we have a promising argument for why consciousness as phenomenal properties must have causal power.

In addition, it would also seem we can add the premise that phenomenal properties are distinct from the fundamental physical properties that it emerged from, for it seems clear that phenomenal properties are not like the kind of properties we find in fundamental physics, or even in neurology or cognitive psychology. Notice however that if we admit phenomenal properties into our ontology, and they have causal power, and we are not epiphenomenalists, then the physical broadly construed, insofar as it is causally closed, should include them.

We now restate the argument with this added premise:

A. Phenomenal properties exist undeniably.
B. Phenomenal properties are distinct from fundamental physical properties.
C. The eleatic principle is true.
D. Therefore phenomenal properties have causal power distinct from those of fundamental physical properties.

I will briefly explain each premise and present some objections and replies. After going through the premises, I will present the conclusion and then give a picture of what conscious, strongly non-reductive causal power can look like.

Premise A; Phenomenal properties exist undeniably.

One might ask; Why is it obvious that phenomenal properties and quaia exist? Remember that qualia are defined as sensory experience or "raw feels". For example the experience of having a toothache or tasting chocolate. Anyone knows from first person experience that these things are very real. Despite this undeniable reality, some philosophers have proposed that these experiences do not really exist after all. For example, Dennet (1988) argues that when we try to find the features of qualia, namely intrinsic private experience, we find nothing that fits the bill (Dennet, 1988). However this claim seems strange. Imagine someone being tortured. It does not help at all to tell him that the pain he feels is really just an illusion or reducible to his screaming or the firing of c-fibres in his brain. Another problem with this eliminativism, is that when it appeals to there being an illusion, i.e. a distinction between the fact that something seems a certain way, and that it is a certain way, this distinction is always relative to a "seeming so" - which arguably presupposes the phenomenal aspect. Finally, given such incredibly reductionist views of consciousness, we should get in return a plausible explanation of qualia. Incredible claims should require incredible evidence. Instead, we only get non-quantitative, psychological explanations couched in vague terms, only hinting at a proper, scientific explanation. In light of this lack of evidence, such a view might also be morally questionable to the extent that it denies the deep reality of suffering without solid proof. Therefore, I believe such views should only be discussed in an exploratory way, and not to be taken to be true unless one is absolutely certain that it is true.

A problem with any kind of reduction of qualia, is that the experience, or qualia itself is absolutely certain, and does not seem to be reducible in any obvious way. As Descartes famously pointed out, it is possible to doubt the existence of everything except your own most intimate experience. While it is not clear that Descartes was talking about qualia as we do now, as this would clearly be anachronistic, it is important to note that by the same token, Descartes did not use the word "thought" like we do now. For instance, in the Geometric Exposition following the second set of replies to the Meditations on First Philosophy (1641) Descartes defines thought in the following way:

"***Thought****. I use this term to include everything that is within us in such a way that we are immediately aware [conscii] of it. Thus all the operations of the will, the intellect, the imagination and the senses are thoughts*" (Descartes 1641 in Jorgensen 2020).

In this way it seems clear, and common sense would agree, that what we cannot doubt are all that we are immediately aware of, not just thoughts in the modern sense, but also the will, the imagination and the sensible aspects of our intimate experience. I also point this out to show that both the "will" and the "sensible" for Descartes is included in the term "thought". In this quote we see that his use of the word "thought" crucially includes everything within us in such a way that we are *immediately aware* of it. Below we will similarly use the concept of the "epistemic self" to talk of the subject stripped of all but immediate experience.

Having this view on what is "indubitable" means that we can, to some extent, isolate this domain that we are immediately aware of, whether we are philosophically able to define it in any further detail or not. If one has the intuition that this domain is the epistemically most certain of all things (even before we have any reflections, ideas or write philosophy), then an anti-reductionist view may seem more plausible. And if one is committed to a broadly physicalist theory, then qualia as a fusion-emergent phenomena that to some extent eliminates some of the parts it fused from and has thereby gained causal power in its own right, seems like a promising way to go forward in understanding the phenomena of consciousness.

Premise B; Phenomenal properties are different from fundamental physical properties

What I mean with phenomenal properties being different from fundamental physical properties, is that any given phenomenal property intuitively possesses different essential characteristics than the fundamental physical properties we learn about in physics.

For instance, a phenomenal property is intuitively more informative; for instance, imagine seeing a vast landscape, and notice how this experience - one among perhaps trillions of experiences one might have at any given moment - cannot compare in informational content to any fundamental force or particle known to fundamental physics as it stands. A picture is worth a thousand words, and probably a lot more in terms of pure informational content. In contrast, elementary particles and forces, reduced to their fundamental instances and most compressed states, require only minimal amounts of information to specify and compute. If we could reduce the informational content of our experience to the information that is in the parts (or "pixels") of our conscious field, then there would be no problem, and the situation would be comparable. However, if our experience is an irreducible informational whole (similar to what happened in the quantum fusion case), then we have to admit that there is a difference between the low-level fundamental physical properties and the emergent property.

Another example is the direct knowledge intrinsic to the phenomenal property. As for instance seeing the phenomenal relations between things in the landscape, portrayed in a certain way. The objective relations between things, whatever that may be, is not what concerns us here. Rather, we are thinking of the precise phenomenal relations between the phenomenal objects, just as we might see the subjective relationship between a road, a green field, and a house in a dream. Notice that even in a dream, we don't just see for instance a house and a road in the abstract, but are presented with them in our imagination/dream sensation as being oriented relative to the other in a precise *seeming* distance away, and so on. And even just focusing on the house it has many parts that are related to each other in precise ways. This is a characteristic of the phenomenal property that cannot be said to belong to any of the related components of the experience, and similarly cannot intuitively be said to belong to any familiar fundamental physical property.

If fundamental physical properties consist in properties that are essentially independent of their relations, it is not clear that they can account for phenomenal properties that essentially consist in the relationality of their parts. However, we may say that quantum mechanics exactly holds that fundamental properties do consist essentially in their relations to each other. In other words, the non-separability of quantum states can be the general phenomena, and the separability may be the special case. However, we cannot underestimate this shift, as it would mean that reduction in the general case might be impossible, if it is not possible to ever isolate the constituent properties. If the fundamental property of any one thing depends on the context of all other things, then reduction becomes a very tricky business. And it is clearly the case, even in quantum mechanics, that there is a difference between separable and non-separable states, and it is fair to say that the classical conception of physics largely assumes separability. Framed in this way, one might say that fusion emergence is exactly a view that takes the non-separability of certain states, in certain situations, as giving rise to fusion emergent phenomena that are generally not well understood.

Thus, in one sense it seems quite clear that this premise holds, to the extent that we assume as given the classical understanding of fundamental physics, where states are separable. However, as we saw in the last chapter on fusion emergence, arguing for an expanded understanding of fundamental quantum physics, enables us to expand what it means to be a fundamental physical property, to such an extent that it is conceivable that phenomenal properties just are fundamental physical properties. This is why ultimately, this is the premise that we reject in this thesis. That is, if qualia is a fundamental fusion emergent property, then it can be understood as a fundamental physical property.

Premise C; The eleatic principle is true.

The eleatic principle has intuitive appeal and is an important principle in a physicalist ontology. However there is much disagreement about whether or not it is true.

One objection to the eleatic principle is that epiphenomenalism is possible. As we pointed out above, epiphenomenalism faces the “absurdity” and “self-stultifying” objection. Again, this is not to say that epiphenomenalism is inconceivable, or that someone might find ways around these problems, it just means that it is a position with serious intuitive, practical and

theoretical challenges. Unless we are forced into this position by other considerations, it naturally makes sense to challenge it here. It is also not a position that is consistent with the position I want to explore in this thesis. I will not go into depth on the pros and cons of epiphenomenalism beyond what I did back in Chapter 3 in this thesis.

Conclusion D; Phenomenal properties have causal power

Given all these premises, we now have the conclusion that phenomenal properties have causal powers. As we have seen this conflicts with the causal closure of the physical, unless we expand what it means to be a physical property to include phenomenal properties.

Given that phenomenal properties have causal power, a question still remains. Why should we expect that phenomenal properties or consciousness generally has the kind of causal power that is useful to the agent?

To investigate this question, let's take a look at the evolutionary argument for phenomenal powers. This argument, as we will see, finds on independent grounds reasons to suspect that our phenomenal properties, or consciousness, has causal power that is useful to the agent. We will look into this question in the next section.

Phenomenal powers

One philosopher who argues for the usefulness of non-reductive mental powers is Hedda Hassel Mørck. In her paper "The Evolutionary Argument for Phenomenal Powers" (Mørck 2018, p. 16) Hedda argues for the existence of phenomenal powers that act as irreducible mental causes. "*The phenomenal powers view is the view that qualia and other phenomenal properties are intrinsically powerful, which is to say that they produce or bring about their effects, or make them happen, in virtue of their intrinsic character alone*" (Mørck, 2018, p. 16). What does this mean? Mainly that phenomenal properties can have causal influence on the world. For instance, when I drive my car and see the traffic light change from green to red, it can be argued that a phenomenal property, the experienced quality of redness, has a causal influence on me stopping my car. Similarly when I feel a pain as I burn my finger, it

can be argued that the pain has a causal influence on me withdrawing my finger.

And so, if the phenomenal property itself is not epiphenomenal, it would seem that there is something somewhat complex psychologically that has causality and cannot be reduced to parts. Mørch uses the evolutionary argument to attempt to show that phenomenal powers exist and that epiphenomenalism is false. A version of the evolutionary argument was first given by William James, who pointed out that experiences of pleasure are connected with safety and survival while pain is linked with harm and death. If phenomenal properties had no causal power, there would be no good reason why animals should not feel pleasure when starving, instead of the pain of hunger for example (Mørck, 2018, p. 2).

Karl Popper also argued for the evolutionary usefulness of phenomenal properties and presented it as the following argument:

1. Phenomenal properties evolved.
2. If epiphenomenalism is true, phenomenal properties are useless.
3. Useless features do not evolve.
4. Therefore, epiphenomenalism is false (Mørck, 2018, p. 3).

The main weakness of this argument, as Mørch points out, is that its third premise is false. It is a known fact that useless features do evolve. Useless features are sometimes correlated with useful, adaptive features as by-products, also called "spandrels" (Mørck, 2018, p. 3).

It is not impossible that phenomenal properties could have evolved as spandrels. However the basic phenomenal properties of pleasure and pain are hard to explain away as spandrels. For instance, if pain is a spandrel, why does it only happen in connection with tissue damage or destruction of the body? If pain was a spandrel we could just as well expect it to appear in connection to other physical processes than tissue damage. Therefore pain cannot easily be explained away as a spandrel (Mørck, 2018, p. 3).

It could be argued that the pain is just a byproduct of avoidance behaviour, however it looks like pain often comes before the avoidance behavior, actually causing it. This suggests that pain itself has causal power in virtue of its phenomenal character alone. *"The kind of intelligible connection that exists between phenomenal qualities and phenomenal powers*

therefore seems like a kind of connection that is not revealed by physics or the physical sciences, but rather only by first-person experience" (Mørck, 2018, p. 29). As Mørch points out, the connection between pain and its powers is mainly a first person experience.

However, is it possible to connect the seemingly true first-person insight that phenomenal powers exist, with third person physics or the physical sciences? In particular, can we use the physical possibility of fusion emergence to explain phenomenal powers? That is what we will investigate further in the next section on qualia as fusion causation.

Qualia as fusion causation

In chapter four we saw that we have reasons to believe that fusion emergence is possible and earlier in this chapter I argued that there are reasons to believe that qualia, or phenomenal properties, have irreducible causal power. Now we will combine these two insights and ask whether or not it is possible that the causal power of phenomenal properties are fusion emergent? If this is possible, then we have shown how phenomenal powers might be possible, at least in principle, while respecting a broadly physicalist ontology.

What fusion emergence first of all shows us is that downward causation is not only possible, but also compatible with a broadly scientific explanation. As Aharonov argued, by suitably balancing the loss for the physical properties that fuse, with the gain of the emergent phenomenal property that gets created, scientific demand for closure and structure can be respected in a quantitative way.

However, we only have examples so far of fusion emergence happening on the level of particles, then could it also happen at the level of mind? The results discussed so far certainly indicate the possibility, although there is not much evidence for these kinds of weird quantum effects at the psychological levels of explanation. In any case, if mental events like qualia can be seen as a *possible* case of fusion emergence, this shows that at least it is *possible* to have non-reductive, conscious mental causation without overdetermination.

If we follow the logic of fusion emergence, the qualia, or phenomenal properties, arise as a fusion of parts that goes into it, such that it becomes a cause of action in its own right. In Aharnov's article, he found that sometimes higher-level correlations in quantum mechanics can not be deduced from lower-level correlations, while the reverse would hold true.

For phenomenal properties then, we might imagine that there are correlations that hold at the phenomenal level, that do not hold for the parts that fused into it. In the sense that phenomenal properties are "what it is like" properties, we can speculate that there is something informative, for instance, about a "what it is like" property, that is not accessible from lower-level properties. In the quantum case, the metaphor was that one had to go through the high level property in order to get information about the parts, that one could not get by going through each part individually. Similarly here, we may have to "go through" the phenomenal property to get certain kinds of information about lower levels that we cannot get otherwise.

The logic of fusion emergence also provides the possibility for downward causation. Because the information available to a higher-level actor is greater than the information available to the lower-level actors, it follows that the higher-level actor may have different powers of causal selection in a given instance. For example, in the quantum case instance, from the vantage point of the higher-level correlation one might have extra information that enables one to choose to measure electron spin in the x-direction rather than the z-direction for some overall purpose (Ellis, 2018, p. 11662). It is a complex question how this happens in detail, and it may be that the causal power of qualia would not be strong enough to cause useful behavior in practise. However as we have said, it is enough to show that this is possible in order to argue that free will is possible. And that of course is mainly what I aim for here.

Objections and replies to the qualia fusion causation view

From quantum to qualia fusion?

The first objection one might have to this possibility, is that I have only given examples from the quantum realm with regards to fusion emergence. There are two ways to reply to this. The

first option is to say that since quantum fusion emergence is a scientific example of fusion emergence, having shown that it occurs in one place, it may be reasonable to expect it to also happen outside the scope of the quantum realm. At least there is no obstacle to look for the phenomena once it is recognized as possible.

This would mean that fusion emergence could be a widely overlooked phenomena, that at its core has nothing to do with quantum fusion emergence, but that the quantum case is an occasion to recognize it generally. On this view, it could even be that a higher-level, quantitative formal theory of consciousness, such as Tononi's integrated information theory view, may be analyzed in terms of fusion emergence, because the information in the whole adds up to more than the sum of the parts, without having to appeal to any fundamental quantum effects. It would take us too far afield to discuss alternatives like these in detail, but the general idea is that fusion emergence may occur in ways that are more general than we currently recognize.

The second option is to try to carve out the possibility that qualia fusion might be an example of quantum fusion at a higher level, that is perhaps beyond the scope of our current science. This would probably sound like new age svada for most physicists. However, even if most processes in nature approximate classical physics, it is not unreasonable to assume that nature makes heavy use of processes that depend on the non-classical aspects of quantum mechanics to be more efficient. While still a fringe view, there is growing support within philosophy for the view that quantum mechanical effects may play a role in explaining aspects of the mental (Stapp, 2008, Chalmers, 2006). There is also growing evidence that certain biological systems use quantum mechanics. In photosynthesis, for example, quantum effects help plants turn sunlight into fuel more effectively (Sarovar et al. 2010 p. 462). Scientists have also proposed that migratory birds have a "quantum compass" enabling them to exploit Earth's magnetic fields for navigation (Hamish et al. 2016, p. 4634). We see that the field of quantum biology is growing rapidly.

From conscious experience to agent causation?

Another objection to the freedom of phenomenal properties is that even though my qualia in itself can have causal power, this does not necessarily give me any freedom. For instance, I might agree that the feeling of hunger can cause my body to walk to the kitchen and make a

sandwich, but as an agent, hunger (in the phenomenal sense) is just one of many phenomenal properties I may have at any given moment, and it does not seem to necessarily follow from the fact that hunger causes my movement, to the fact that it is I, the agent, that is causing my movement. In other words, it is not sufficient for a quale to cause movement to say that I am causing that movement. In addition, as an agent it seems plausible to suppose that complex and difficult choices require more than just a few phenomenal properties. Thus, the problem still remains of showing how this causation has anything to do with freedom in the sense we are looking for, namely the freedom of the agent.

It is true that most of our actions are not chosen with any great awareness, willpower or through any extensive reasoning by the agent. Our survival instinct, habitual actions, routines, etc. usually take care of these things more or less by themselves, without the need for a lot of conscious choices being made or a number of possible actions being rationally considered. However, it still seems that something beyond appealing to the causality of phenomenal properties is required to explain the freedom that is available to the agent. In general it seems that the agent's free will involves something more than the qualia I experience. This "more" may be fleshed out in many different ways, but since we have argued for the strongly emergent power of phenomenal properties, we here focus on giving a brief sketch of how phenomenal properties help enable the agent to act freely. However, by briefly outlining such an account, and focusing on how phenomenal properties as fusion emergent might be the "missing piece", we go some distance towards showing how such an account is possible. In the next chapter I will look into what it means to be a "free agent", and I will try to draw the relevant links between downward causation, consciousness, agent causality and free will.

Chapter 6: The freedom of a conscious agent

What is an agent?

The term “agent”, as used in this connection, is defined as “one who acts or exerts power, or something that produces an effect, an active or efficient cause” (Merriam Webster dictionary, 2020). In addition, I take agents to be essentially conscious. As conscious agents, we are aware that we exist. The problem of personal identity generally deals with how persons are defined, and what we identify with as acting, thinking, feeling beings. In one sense personal identity can be about what makes someone a person or the essence of personhood in general. In another sense, personal identity can be about what distinguishes one person from other persons, and what accounts for the identity of a person across time (lifetime).

Also, being an agent typically involves the notion of intentional action, which means that the action is somehow directed towards achieving some end, in the sense that the action is about something. Thus the notion of an agent that we will investigate here, involves both the notions of personal identity and intentional action. I first give a short explanation of personal identity across time, then briefly go into intentional action. Finally I will give my own account of what a free agent is, and why the consciousness of the free agent, as fusion emergent qualia, ensures the possibility of a strongly emergent free-willed agent.

Personal identity

What sort of things are we metaphysically speaking? What are our fundamental properties and are we composed entirely of matter or are we partly immaterial? Are we spatially extended and if so where are our spatial boundaries? Are we independent substances or aspects of something else? These are difficult questions and different thinkers suggest different answers. Some suggest that we are biological organisms, or parts of biological organisms, like brains. Others suggest that we are partly immaterial souls, or collections of

mental states and events, some have even suggested that we do not really exist at all. There seems to be no consensus or dominating view on this problem (Olson, 2019).

Personal identity over time

Any theory about personal identity and agency, must take into account what it means for a person or an agent to persist through time. What determines which past or future being is you? Let's say you look at old photos from your childhood and you say "that's me". What makes you that particular child and not another? How can you as a child and you as an adult be numerically identical? The physical body changes and even though you might look vaguely similar to yourself as a child, all the cells in your body have changed many times from then to now (Olson, 2019).

The most popular answer to this problem is the "psychological-continuity view". This view says that your persistence consists in some psychological relation between the child version of you and the adult version of you. The same mental features like beliefs, memories, preferences and the capacity for rational thought are inherited by all the versions of you that persist through time (Olson, 2019).

In this thesis we are mainly concerned with how an agent at any particular time, for example at time X, can make a free choice. This idea of an agent with causal power to choose freely, can fit with whatever theory of personal identity one might adhere to. For our purposes here we cannot go into any specific theory of what a person is, or whether or not this person can persist over time in order to establish a causally efficacious agent. We only assume here that there is an agent at any given time X that is able to choose, or bring about an effect.

Intentional action

According to the agent-causal view I promote in this thesis, agency cannot be reduced to mere events, because only subjects with minds can have the ability to act. It is me as an agent

who does something and is the source of my action, it is not something that happens in me (Moya, 1990, p. 48).

Intentional actions consist in aims, beliefs, plans, rules, future and immediate intentions (Moya, 1990, p. 58). Some intentions seem to be limited to minds and require a certain level of rationality and coherence, complex actions therefore are not isolated items, they can only make sense in the broader context of rational minds (Moya, 1990,p. 61). However, I don't think animals can't have intentional actions. From a common sense view, animals have phenomenal properties and intentional agency, but they have a much weaker form of this capacity, for instance, they can't reflectively consider options and make decisions about the far future. This gives us two senses of agency, weak and strong.

Weak and strong agency

We can separate between full blooded agency and a weaker, minimal sense of action. This latter sense are expressions of immediate intentions, for example, the intention to change our sitting position when we feel uncomfortable, or the intention to drink from the glass I have before me. These intentions are minimal, in that they do not involve any commitment to future actions, but are spontaneous responses to immediate wishes and needs. Even small children and animals can show intentional behavior in this weaker sense of agency.

The stronger sense of agency however is only possible for more mature human brains, because it involves a robust sense of commitment to an action and the ability to take responsibility for it. For example the commitment to marry someone in two years time. This conscious commitment and responsibility for her actions are not only assigned to the agent, but consciously chosen and taken on by her. This second aspect is not found in the intentions of animals, they only show scattered features of intentionality and have no sense of norms and commitment to future actions (Moya, 1990, p. 67). In order to have future intentions, the agent must both understand the concept of time, and possess the means to represent and refer to what is not immediately present to the senses, which is to master abstract thinking (Moya, 1990, p. 131). To the extent that we want to argue that animals have free will, this will only be a limited sense of freedom in comparison with the freedom adult humans have available.

What is a free agent?

For our purposes here I suggest a minimal definition of an agent which I call the epistemic self. The epistemic self is related to qualia and overall consciousness. The epistemic self consists in all the things we can know indubitably about ourselves at any given moment of experience, the prime example being that we exist. Animals also have an epistemic self, even if they cannot articulate any of the knowledge they possess by virtue of their phenomenal properties, for example they can phenomenally feel their body being in a certain position.

I mention animals here, because some philosophers tend to over-intellectualize the "knowledge" we have in virtue of phenomenal properties. Since we assume that animals and infants also have phenomenal properties, we must grant them this kind of immediate epistemic knowledge. Descartes, although there are conflicting opinions on this, also seems to subscribe to a similar view, not with regards to animals but infants specifically:

"*In view of this I do not doubt that the mind begins to think as soon as it is implanted in the body of an infant, and that it is immediately aware [conscius] of its thoughts, even though it does not remember this afterwards because the impressions of these thoughts do not remain in the memory*" (Descartes in Jorgensen 2020).

Thus the "cogito" in Descartes, is arguably so non-intellectual that it can be very difficult to grasp its essence in its simplicity, due to our tendency to over-conceptualize our experience. It is also plausible to suppose that even adult humans do not retain the memory of our present experience in any detail for any length of time, as our attention shifts every fraction of a second, and so discards information that is no longer useful to the agent as a whole. The minimalist epistemic self is extremely minimal in this sense, since any decision could be based on the moment to moment fusion of experience. For instance, when in a flow state a cat is chasing a bird, the cat is using the information captured second by second in order to make optimal decisions about its movements and hunting strategy.

In general the focus on Descartes' argument as only establishing the I's indubitable existence ignores the wealth of everyday things that an epistemic self knows directly, in the moment of presence. For example the taste of coffee when we are tasting it, the way our living room

looks to us when we see it, the way our body and arms feels to be oriented in a certain way, the particularity of the sounds we hear as we notice them, and so on. Because these things appear so basic and indubitable at the moment we experience them, it might be said that having conscious experiences of one kind or another is essential to being an agent. If there was nothing it was like to be an agent, there would be no intrinsic subjectivity, so it could just as well be a zombie agent. I won't argue here that this is impossible in principle, rather it seems that in our world the kind of essential freedom and causal power we have as agents is partly a result of our consciousness, or at any rate that is what I am trying to argue here.

The epistemic self and qualia-space

While the things within the epistemic self at any moment are linked to particular experiences, and particular phenomenal qualities, the epistemic self as a whole does not have to be linked to any particular experience, as it seems to be like an overall medium in which all indubitable experiences happen. For example I do not identify with any single part of my body, as it feels obvious that I am not only my arm or my foot or my head, however the experience can include many of these aspects at the same time.

This means that the epistemic self is the whole within which our perceptions and our actions occur. Somehow, we can be aware of the taste of coffee, the positioning of our body, and a beautiful piece of music, all at the same time. Psychologically, it is clear that many different areas of the brain, for example our auditory cortex, our visual cortex, and somatosensory cortex, are all connecting in such experiences, coming together in what some call a "global workspace" or in Tononi's more technical, formal sense a "qualia-space" (Balduzzi and Tononi, 2009, p. 1). Crucially, Tononi's idea is that such a qualia-space is always more than the sum of its parts, since neither, for instance, the coffee cup or the muffin before me can give me the effect of being able to see their relationship, or ability to select between them.

In scientific terms we can think of a qualia-space as the integrated connection between all the qualia happening in a single brain or system. The system can be more or less complex, as in the case with animals or small children. The overall freedom of the agent seems to depend on the overall maturity and complexity of her qualia-space. To take a simple example, if I am

color blind, not just visually but also with a color-blind imagination, then I will not be able to imagine colors, or paint what I imagine, or navigate according to colors in the world; on the other hand, if I have some artistic talent, perhaps being a tetrachromat, and additionally have studied the art of painting, and developed my capacity to distinguish different colors, I may then have the freedom to both perceive, imagine, and create a large range of paintings, in addition to navigate the world in a different way than most people.

A more embodied example is the difference between the bodily freedom of a dancer and a non-dancer; the dancer can express a multitude of possible expressions, while the non-dancers movements will all fall within a relatively narrow range, even as they try to do something random. In a similar sense to the color example, the dancer can perceive, imagine and move in ways that may simply be imperceptible, unimaginable and un-actionable for the normal person. Naturally as philosophers, we may consider the freedom of thought to be supreme, but as we know, even that must be developed in such a way that we can intuit the "chessboard" of philosophical moves and navigate them with a freedom that others may not be able to do.

One may object to this being a consequence of some "qualia-space" but we simply mean here the sum of phenomenal properties available to the epistemic self. The epistemic self in turn was simply defined as our immediate direct experience that is present in any given moment. Of course, to quantitatively establish this with any mathematical or scientific precision we would need to use a theory like the one Tononi and his colleagues are promoting (Oizumi, Albantakis and Tononi, 2014, p. 13), or some other theory that might do similar work in classifying the level of integration required for qualia to arise in any given moment.

We can therefore roughly say that free will comes in degrees. As we grow up the connections in the qualia-space system also grows and develops, until we as adults become relatively mature and free agents, with greater control and responsibility for our actions. By emphasizing the idea that free will comes in degrees, it will also be more compatible with a scientific view, where the fusion emergence will need to take into account the development of the brain and the appropriate levels of fusion that can occur in any given instance. All indications point to fusion emergence as being a delicate affair in practise, difficult to arrange and costly to sustain.

This is not to disregard that there may be jumps in freedom, that are not just gradual, as may have occurred in the transition from animals to humans for instance, in relation to gaining some capacity for language and reasoning. Another example may perhaps be if humans to some extent or other can reach and sustain certain peak experiences that may also imply a jump in freedom. Those kinds of considerations are outside the scope of this essay however, since the focus here is on establishing the most minimal conception of agency.

It is relatively straightforward to see that the agent needs to have a unified perspective in particular instances, for example to be able to see the structure of the room with certain people and furniture in it. It is perhaps less clear that some kind of unity is required for action. However, in order to select a course of action, at least in particular instances, it is similarly required that the agent have a range of alternatives to select from, that are considered in a unified perspective. This is why the epistemic self is relevant to the freedom of the agent in general, because it is to the extent that we are epistemically aware of the fact that we have a certain choice, that we are free agents. We all have had the experience of making some choices less consciously than others, often from some impulse or other, that has happened without the light of having a clear-headed consideration of the whole situation. The freedom of action that comes from taking into account the whole situation with all the available alternatives, when we are really awake, is clearly something displaying a higher freedom, and that's why we also have to allow for degrees of freedom.

The fusion emergent qualia space and the epistemic self

Earlier I argued that mental causation must be a fusion emergent, downward causal power. The qualia-space of an agent should therefore be a fusion emergent and integrated entity, meaning that it exists independently of its underlying micro-physical properties. This does not mean that it is not ultimately physical, just that it exists independently, since in the process of fusion, the constituents give up some of their existence to the fusion emergent property. This causal power must be closely linked with the physical world. The fusion of small elements on the micro-levels of physics, appears to be able to create new, independent macro-causal power on a high physical level in brains or other complex physical systems. It is this macro-causal power I wish to link with the mental, i.e. as being a phenomenal power.

In the first instance, I held that qualia fusion emergence could explain the emergence of phenomenal properties. Phenomenal properties are not sufficient however to explain agency, to the extent that we view these properties as components or parts that exist within the epistemic self. We therefore introduced the notion of a qualia-space, to explain that phenomenal properties are integrated together in a single brain or system. Now the missing piece consists in connecting the qualia space to the epistemic self.

The most natural way to understand the relation between qualia-space and the epistemic self is to say that qualia-space simply *is* the epistemic self. One may have other ways to understand this relation, but this is clearly the simplest. The unity and freedom of the epistemic self, is therefore nothing but the unity and freedom of the qualia-space. The qualia-space in turn emerges from phenomenal properties just as phenomenal properties emerge from fundamental physics, by fusion emergence. We thus have an emergent, macro-causal, phenomenal power (qualia-space) that can be identified with the epistemic self.

Remember that the epistemic self is just one aspect of our personal identity, a very thin conception that only concerns the indubitable experiences we have from moment to moment. However, we argue here that even this thin aspect has a capacity for unity and freedom that strongly emerges and gives rise to many strange phenomenal powers that are irreducible to any lower level explanations. While this freedom is not the strong freedom of developed humans, but rather the humble freedom that we to a large extent share with animals, it can nevertheless serve as the basis for many of our higher functions, and may lay the groundwork to flesh out a stronger freedom that only applies to humans.

No matter what circumstances or surroundings you might find yourself in, whether you are the president of the United States, a poor beggar in the streets or a monk in the middle ages, or simply some animal looking to explore the landscape, the amount of development of your qulia-space can give you more or less freedom to choose your actions. Of course the overall possibilities you have for available actions, will be vastly different depending on your biology, neurology, age, skillset, education, etc.

Nevertheless, the agent, insofar as he is an epistemic self, will be enabled by the fusion-emergent qualia-space in any given moment of experience, to have some freedom over and beyond the elements that enter into his particular circumstances. For instance, if within our

epistemic self (qualia-space) there are five different impulses, to eat, explore, sleep, or socialize, the decision between them is made on the level that integrates all of these impulses into a single irreducible decision. Or if we are painting a picture, and we have many different pictures, colors and forms in memory and imagination, we can now produce something altogether unique, that goes beyond all the influences that gave rise to the painting.

Now, let us return to the question: what does consciousness have to do with free will? Cannot free will exist without consciousness?

Why are free agents conscious?

For most people it seems obvious that free agents must be conscious. We see this sentiment clearly expressed by Merlin Donald in this quote; "*There is clearly no free will, in any meaningful sense, outside conscious awareness*" (Donald. M, 2010, p. 9.). Many modern cognitive scientists however have argued that consciousness, along with freely chosen action, is nothing more than a comforting illusion. The claim is that even though we clearly experience ourselves as being free and fully aware when we make a decision to act in a certain way, empirical evidence, like Libet's studies mentioned earlier, arguably seem to indicate that it may be an illusion. The motivation to dismiss consciousness as a comforting illusion may also be due to how terrifyingly difficult the problem is to explain within a broadly physicalist framework. It is generally difficult for any new theory to gain ground when the terminology and the development of the theory is in its infancy, versus an older theory that has several adherents and has been battle-tested and hardened against critique. However, this difficulty should not be an argument against working on theories in their infancy, otherwise there would never evolve any new theories.

The idea that consciousness is an illusion was first formulated in the behaviouristic movement in psychology, that only focuses on observable behaviour and disregards mental activities. The main idea here is that we are unconscious, deluded automatons who only think we have some kind of intrinsic unified subjectivity, but that it is really an illusion (Donald, 2010, p. 10.). By the success of behaviorism, in the sense of focusing science on questions that we can be answered behavioristically without invoking any unexplained phenomena, or

the focus on finding neural correlates of consciousness instead of seeking an explanation of why they correlate just so, the behaviouristic “no nonsense” approach to some extent enabled progress and still has a strong hold on modern thought.

Behaviourism seems to have been a strong influence towards not seriously considering the possibility for free will. However as Merlin Donald points out, the arguments showing strong evidence for unconscious influences on the mental are not fatal for free will in their own right, because they are inductive arguments. The problem for such arguments is that they only show that unconscious influence is important, they cannot in themselves disprove the possibility of free will (Donald, 2010, p. 10.). This shows that the possibility for conscious free action is still open.

The Robot Example

According to event-causal accounts, free will is not caused by the conscious agent, but by events happening in the brain. In this picture consciousness might not be necessary for free will. Let us imagine a very clever robot and a human agent, both choosing between various actions with equal success. The only difference is that the human has a phenomenal consciousness and the robot does not. This is not to say that artificial consciousness is impossible, we just assume for the sake of argument that robot means non-conscious robot. Why can't the robot have free will? I want to argue that “of course!” the robot cannot have free will - it just acts automatically! However humans often act automatically too, skilled drivers need not be conscious of every detail of the whole driving process, yet we still want to say that they are acting freely. We could say that the robot does not have any goals or intentions, but it is easy to program the robot to prefer certain outcomes or reach a goal.

It seems that the crucial difference between the robot and the human may be that the human has an epistemic self. Which means, according to the eleatic principle, that it has an irreducible causal power, as we argued with the phenomenal properties argument. The epistemic self construed as being fusion emergent, means that it is a macro-structure that is more than the sum of its micro-parts. In contrast, it seems that no such line of reasoning is available for the causal power of the robot, since it has no strongly emergent phenomenal properties, as so has no costly unity associated with it. Thus it seems it can still in principle

be reduced to the bottom-up causal power of complex aggregates of physical particles acting together, and so does not need to exist in its own right according to the eleatic principle.

The fusion emergence of the epistemic self means that the human can take the whole situation (as it is presented to the phenomenal subject) into account directly, as existing indubitably within the epistemic self, before choosing an action. The robot on the other hand can only take into account the information indirectly, through an aggregate process. However, if we create artificial intelligence that could similarly act as a unified, non-reducible causal whole, we could expect it to have become conscious in the process. By that point it is reasonable to assume, at least in these nomological circumstances, that it would have to go through some fusion emergence process, and so it would have become real at a level that is more than the sum of its parts. It is likely that, in these nomological circumstances, any high-level irreducible complexity of the type required for artificial intelligence, would be qualia fusion emergent, and expressed as consciousness, just as it seems to be in us. This question is ultimately a scientific one however, and there is no need to speculate on it here.

To illustrate the difference between irreducible and reducible causes from a different point of view, we can use the example of a mother holding her baby and a robot holding a baby. They are in the same situation and their arms get damaged by a bullet. Some receptors in the brain of the mother and some sort of wiring in the system of the robot, informs them both that they have to let go of the baby to avoid pain or getting their arm seriously hurt. The robot lets go of the child automatically, but the mother does not. Why do they act differently? Because the mother is able to take into account the whole situation and judge between many different possible outcomes, and in that light makes a choice of not letting go of the baby. Perhaps she is followed by someone dangerous, or there is an angry dog on the ground, maybe the child is hurt, maybe she is outside in the winter and has to get inside before the child gets too cold. There are countless situations where the pain of the mother's arms will not be taken into account. The robot in contrast, would only sum up the respective causes and act according to a “sum” of all those influences, with no irreducible causation occurring. In other words, the Robot would use a general rule of combination, or general procedure, for all combinations, whereas the mother can make a unique irreducible high-level decision taking all into account.

We can argue that humans also just act automatically without any conscious engagement, this is true when we engage in activities that we have practiced as a routine. But, even in these situations an agent can consciously intervene in his behaviour at any time. For example when you are automatically driving a car, and suddenly have to stop for a child running into the road, or when a pianist who plays automatically, at some point consciously chooses to modify his well-rehearsed piece (Donald. M, 2010, p. 10.). We always have an option to consciously intervene in our automated actions, unconscious robots do not have this option. This would also help provide a reason why consciousness evolved in humans, because it is useful to evaluate complex situations as a whole, and respond to the whole situation jointly, rather than just summing up individual rules or methods.

In addition, if our perception for instance is just a sum of an aggregate of information, it is not clear why it wouldn't just be epiphenomenal. But here of course we reject epiphenomenalism, and then it seems clear it must do something beyond the aggregated parts that gave rise to it. In contrast to unconscious processes, the epistemic self is able to choose possible actions and outcomes that are highly unlikely and yet extremely precise, all things considered. This might be because we have access to this higher level fusion emergence and downward causation, that cannot simply be broken down to the lower-level causal happenings on the atomic or molecular levels of physics.

The freedom of a conscious agent

I have argued that the freedom of a conscious agent is made possible by her direct experience of the world, the epistemic self. As an epistemic self, the agent is able to see and act in the moment, giving rise to a minimal notion of a free agent. Taking the phenomenal properties and evolutionary argument as pointing to strongly emergent phenomenal properties, we took the phenomenal property as the basis for the qualia-space, or the epistemic agent, within which it makes sense to define a minimal notion of the freedom of the agent.

As such, the free agent becomes a self-determining cause of her own actions, insofar as the epistemic agent, as having a joint capacity for perception and action, cannot be reduced to anything else. In this sense, our minimal freedom is very much like our perception, in that the

unity of perception “overflows” the elements that go into it, the unity of action “overflows” the possible actions that went into determining the actual action.

Just like perception is more than the sum of its inputs, if we understand it as an irreducible whole, the action is more than the sum of its possible alternatives, if we understand it in light of the epistemic self as a whole. When we create a painting, we see why these aspects are so relatable, because we must both be able to consider the picture as a whole, over and beyond each part, and be able to consider the next stroke, in light of all the possible strokes we might have painted next, to be able to freely paint some masterpiece that is irreducible to anything that went into the work.

With fusion emergence, we have a plausible way forward that respects broadly physicalist intuitions. There is a balance between the loss of the fused physical properties that go into our epistemic self, with the gain in phenomenal properties at the level of our epistemic self, such that there are no structural inconsistencies for a broadly physicalist framework.

Naturally this is only the briefest sketch of such a view, and I realize that this view may take a long time to develop into maturity. Here I mainly wanted to defend the possibility for free will within a broadly physicalist framework, and with the phenomenal properties argument, together with fusion emergence and a minimal notion of the epistemic self, I believe that this is a position that can be made consistent and coherent in a longer treatment.

Conclusion

In this paper I have attempted to argue for a robust sense of free will by appealing to a kind of agent causation, where the agent is interpreted as an epistemic self, and the causation consists of fusion emergent, conscious mental powers.

I started the discussion about free will by contemplating the different attempts to make sense of free will by various philosophers throughout time. In the first chapter we saw that the three main positions were strong determinism, compatibilism and libertarianism. In the second chapter I choose to defend agent-causal libertarianism, because this is the only position that aims to give the robust sense of freedom needed to explain our intuitive experience of being free to choose our actions and control our physical bodies and surroundings independently of any predetermined happenings.

In the third chapter I gave an account of mental causation and chose to argue for strong emergence, because weak emergence could not give a plausible account of how mental causation can happen independently of its micro-physical realizers. Instead I attempted in chapter four to show that such independence is possible by appealing to fusion emergence. We saw that fusion could give us the radical kind of downward causation needed to explain mental causation. The question still remains open however, if science will be able to give an explanation of how the brain can generate this kind of downward causation.

In chapter five I argued that mental causation must be conscious in order to count as free at all. It makes no sense that we are free agents without even being aware of our freedom. I hold that qualia is the heart of consciousness, and that qualia must be a strongly fusion emergent property with downward causal power. I think this is what free will ultimately must be, if we are to explain the full sense of freedom that we experience as human beings. In chapter five I suggested that the conscious agent is essentially the epistemic self, consisting of a unified qualia-space, able to act freely with downward causal power. This minimal notion of a free agent can further be supplemented with a broader theory of agency and personal identity.

My main motivation in this thesis has been to carve out a possibility for a strong free will, that respects the freedom we experience as real, while still being true to the physicalistic

mindset of modern science. I believe that the epistemic self as characterized by qualia space and emerging through qualia fusion emergence is a promising approach to solve the problems of strongly emergent free will within a broadly physicalist framework.

References

- Aharonov, Y. and Bohm, D. (1959). Significance of Electromagnetic Potentials in the Quantum Theory. *Physical review*, 115(3), p. 485-491. doi: 10.1103/physrev.115.485.

- Aharonov, Y. Cohen, E. & Tollaksen, J. (2018). Completely top–down hierarchical structure in quantum mechanics. *Proceedings of the National Academy of Sciences of the United States of America, 115*, p. 11730 -11735. doi: 10.1073/pnas.1807554115.

- Aharonov, Y. and Casher, A. (1984). Topological Quantum Effects for Neutral Particles. *Physical review letters*, 53(4), p. 319-321. doi: 10.1103/physrevlett.53.319.

- Aharonov, Y. David, Z. and Vaidman, L. (1988) How the result of a measurement of a component of the spin of a spin-*1/2* particle can turn out to be 100. *Physical review letters*, 60(14), p. 1351-1354. doi: https://doi.org/10.1103/PhysRevLett.60.1351.

- Aharonov, Y. et al. (1993) Measurements, errors, and negative kinetic energy. *Physical review. A*, 48(6), p. 4084-4090. doi: 10.1103/physreva.48.4084.

- Aharonov, Y. et al. (2017) The Mathematics of Superoscillations. *Memoirs of the American Mathematical Society*, 247(1174), p. 1-105. doi:10.1090/memo/1174

- Aharonov, Y. et al. (2013) Quantum Cheshire Cats. *New journal of physics*, 15(11), p. 113015. doi:10.1088/1367-2630/15/11/113015.

- Aharonov, Y. and Vaidman, L. (2008). The Two-State Vector Formalism of Quantum Mechanics, in Muga J. G. et al. (eds.) *Time in Quantum Mechanics*, Berlin: Springer, p. 399-447.

- Agent (2020) in *Merriam Webster dictionary*, Available from: https://www.merriam-webster.com/dictionary/agent (Retrieved: September 18. 2020).

- Balduzzi D. and Tononi, G. (2009) Qualia: The Geometry of Integrated Information, *PLoS Computational Biology*, 5(8), p. 1-24. doi: 10.1371/journal.pcbi.1000462

- Chalmers, D. J. (2006). Strong and weak emergence. In Davies P. and Clayton P. (eds.) *The Re-Emergence of Emergence: The Emergentist Hypothesis From Science to Religion.* Oxford: Oxford University Press.

- Clarke, R. and Capes, J. (2017) Incompatibilist (Nondeterministic) Theories of Free Will, in *The Stanford Encyclopedia of Philosophy*, Available from:

https://plato.stanford.edu/archives/spr2017/entries/incompatibilism-theories/ (Retrieved: November 29. 2019).

- Dennet (1988) D. C. Quining Qualia. in A. Marcel A. and Bisiach, E. (eds.) *Consciousness in Modern Science*, Oxford: Oxford University Press. Available from: http://cogprints.org/254/1/quinqual.htm (Retrieved: November 27. 2020).

- Donald, M. (2010) Consciousness and the Freedom to Act, in Baumeister, R. F. Mele, A. R. and Vohs K. D. (eds.) *Free Will and Consciousness: How Might They Work?* New York: Oxford University Press, p. 9-23. Available from: https://oxford-universitypressscholarship-com.ezproxy.uio.no/view/10.1093/acprof:oso/9780195389760.001.0001/acprof-9780195389760. (Retrieved: November 24. 2020).

- Ellis, G. F. R. (2018) Top-down causation and quantum physics. *Proceedings of the National Academy of Sciences of the United States of America,* 115(46) p. 11661-11663. doi: 10.1073/pnas.1816412115.

- Hassel Mørck H. (2018) The evolutionary argument for phenomenal powers. *Philosophical Perspectives,* 31(1), p. 293-316. doi: https://doi.org/10.1111/phpe.12096.

- Hoefer, C. (2016) Causal Determinism, in *The Stanford Encyclopedia of Philosophy*, Available from: https://plato.stanford.edu/archives/spr2016/entries/determinism-causal/ (Retrieved: April 18. 2020).

- Hamish, G. et al. (2010) The quantum needle of the avian magnetic compass. In Klein L. (ed.) *Proceedings of the National Academy of Sciences of the United States of America,* 113(17) p. 4634-4639. doi:https://doi.org/10.1073/pnas.1600341113.

- Hoel E.P. Albantakis, L. and Giulio, T. (2013) Quantifying causal emergence shows that macro can beat micro, *National Academy of Sciences*, 110(49), p. 19790-19795. doi: https://doi.org/10.1073/pnas.1314922110.

- Humphreys, P. (1996) Aspects of Emergence, *Philosophical Topics*, 24(1), p. 53-70. Available from: http://www.jstor.org/stable/43154222.

- Humphreys, P. (2016) *Emergence: A Philosophical Account*, Oxford, Oxford University Press. doi: 10.1093/acprof:oso/9780190620325.001.0001.

- Jorgensen, L. M. (2020) Seventeenth-Century Theories of Consciousness, in *The Stanford Encyclopedia of Philosophy* Available from:

https://plato.stanford.edu/archives/spr2020/entries/consciousness-17th/. (Retrieved: December 13. 2020).

- Kim, J. (2006) Emergence: core ideas and issues, *Syntese, New Perspectives on Reduction and Emergence in Physics, Biology and Psychology*, 151(3). p. 547-559. doi: 10.1007/sll229-006-9025-0.

- Kim, J. (1998) *Mind in a physical world*, Massachusetts: The MIT Press.

- Kim, J. (2005) *Physicalism, or something near enough.* Princeton: Princeton University Press.

- Kirk, R. (2019) Zombies, in *The Stanford Encyclopedia of Philosophy.* Available from: https://plato.stanford.edu/archives/spr2019/entries/zombies/. (Retrieved: November 29. 2020).

- Loewer, B. (2011) From physics to physicalism, in Lepore, E. and Loewer, B. (eds.) *Meaning, mind and matter: Philosophical essays,* Oxford: Oxford University Press. Available from: https://oxford-universitypressscholarship-com.ezproxy.uio.no/view/10.1093/acprof:oso/9780199580781.001.0001/acprof-9780199580781-chapter-14 (Retrieved: October 20. 2020).

- MacBride, F. (2020) Relations, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/archives/win2020/entries/relations/ (Retrieved: November 20. 2020).

- McKenna, M. and Coates, J.D. (2020) Compatibilism, in *The Stanford Encyclopedia of Philosophy.* Available from: https://plato.stanford.edu/archives/spr2020/entries/compatibilism/ (Retrieved: Mai 10. 2020).

- Mičuda, M. *et al.* (2017) Experimental demonstration of a fully inseparable quantum state with nonlocalizable entanglement, *Scientific Reports,*7 p. 1-11 doi: https://doi.org/10.1038/srep45045.

- Moya, C.J. (1990) *The philosophy of action, an introduction*. Cambridge: Polity Press.

- O'Connor, T. and Franklin, C. (2019) Free Will, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/entries/freewill/ (Retrieved: Oktober 1. 2019).

- O'Connor, T. (2020) Emergent Properties, in *The Stanford Encyclopedia of Philosophy*, Available from: https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/ (Retrieved: September 14. 2020).

- O'Connor, T. (2005) Libertarian views: Dualist and Agent-Causal Theories. In, Kane, R. (ed). *The Oxford Handbook of Free Will*. Oxford: Oxford University Press, p. 338-355. doi: 10.1093/oxfordhb/9780195178548.003.0015.

- Olson, E. T (2019) Personal identity, in *The Stanford Encyclopedia of Philosophy*, Available from: https://plato.stanford.edu/archives/fall2019/entries/identity-personal/ (Retrieved: September 19. 2020).

- Oizumi, M. Albantakis, L. and Tononi, G. (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology.* 10(5), p. 1-25. doi: https://doi.org/10.1371/journal.pcbi.1003588

- Pereboom, D. (2004) Is our conception of agent-causation coherent? *Philosophical Topics, Agency.* 32(½), p. 275-286. Available from: https://www.jstor.org/stable/43154439.

- Rigoni D. et al. (2011) Inducing Disbelief in Free Will Alters Brain Correlates of Preconscious Motor Preparation: The Brain Minds Whether We Believe in Free Will or Not. *Psychological Science.* 22(5), p. 613-618. doi: https://doi.org/10.1177/0956797611405680.

- Robb, D. and Heil, J. (2018) Mental Causation, in *The Stanford Encyclopedia of Philosophy.* Available from: https://plato.stanford.edu/archives/fall2018/entries/mental-causation/ (Retrieved: Oktober 15. 2018).

- Robinson, H. (2017) Dualism, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/archives/fall2017/entries/dualism/ (Retrieved: 22.10.2018).

- Robinson, W. (2019) Epiphenomenalism, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/archives/sum2019/entries/epiphenomenalism/ (Retrieved: October 19. 2020).

- Sarovar, M. et al. (2010) Quantum entanglement in photosynthetic light-harvesting complexes. *Nature physics* 6, p. 462–467. doi: https://doi.org/10.1038/nphys1652

- Schlosser, M. (2019) Agency, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/archives/win2019/entries/agency/ (Retrieved: October 3. 2020).

- Searl, J. R. (2010) Consciousness and the Problem of Free Will, in Baumeister, R. F. Mele, A. R. and Vohs K. D. (eds.) *Free Will and Consciousness: How Might They Work?* New York: Oxford University Press, p. 122-134. Available from: https://oxford-universitypressscholarship-com.ezproxy.uio.no/view/10.1093/acprof:oso/9780195389760.001.0001/acprof-9780195389760. (Retrieved: November 24. 2020).

- Stapp, H. P. (2008). Philosophy of Mind and the Problem of Free Will in the Light of Quantum Mechanics,in *Lawrence Berkeley National Laboratory University of California*. Available from: https://escholarship.org/uc/item/3852g8c2#main (Retrieved: August 28. 2020).

- Van Gulick, R. (2018) Consciousness, in *The Stanford Encyclopedia of Philosophy*. Available from: https://plato.stanford.edu/archives/spr2018/entries/consciousness/ (Retrieved: September 2. 2020).

- Walter, S. and Heckmann, H. D (2003) Introduction, in Walter, S. and Heckmann, H. D (eds.) *Physicalism and mental causation, the metaphysics of mind and action.* Exeter: Imprint Academic, p. 3-10.

- Will (2020) in *Merriam Webster dictionary,* Available from: https://www.merriam-webster.com/dictionary/will (Retrieved: September 5. 2020).