Université catholique de Louvain

Faculté de philosophie, arts et lettres

Institut Supérieur de Philosophie– Centre de Science et Sociétés

The Gordian knot of Free will.

The causal efficacy of mental events.

Membres du jury :

Prof. Bernard Feltz, promoteur
Prof. Marcus Missal, promoteur
Prof. Peter Verdee, lecteur
Prof. Alexandre Guay, lecteur et secrétaire.
Prof. Helen Beebee, lecteur extérieur
Prof. Nathalie Frogneux, président

Thèse réalisée par
Anna Drozdzewska

en vue de l'obtention du grade de
Docteur en Philosophie

Louvain-la-Neuve
Année académique [2017-2018]

Acknowledgements

Those who have finished a PhD, and probably many that have not, know how demanding and often emotionally draining the process can be. It is also very rewarding if, like me, you had the chance to work on a project you love, on a topic you are truly passionate about, and be able to talk to people that share your enthusiasm. All this however would not be possible without a number of people.

First I would like to thank my two supervisors for giving me a chance to work on this great project: prof. Bernard Feltz, and prof. Marcus Missal, who both took this leap of faith with me, and disregarded my poor knowledge of French. They pushed me when I needed to be pushed and granted me the possibility to present my work to a number of audiences, to test my theories and train my skills. For all that, thank you. Nothing would have happened without the constant support from the team in UCLouvain, which was my academic home for four years. Alexandre, who was always there when I needed him, Peter who sat through my numerous rants about interventionism, always patient and interested, Andrew who read through the papers and gave me plenty of advice and Olivier who always had a sympathetic ear. To all of you, and others in the department a very warm thank you, I would not have been able to do it without you. My gratitude also goes to the administrative team who helped with the bureaucracy. I would like to especially acknowledge Monique, Laetitia and Valeria, all of which were of great help in getting here, lending me a hand in navigating the administrative meanders.

Aside from the working environment, also the support of friends was essential. A thank you for all the lunches and dinners, which always helped to keep my spirits up, goes to the crowd from Leuven. Although some of us moved far, far away we still manage to be of support for one another. Thank you all for the years of friendship, and I hope to further see our children growing up together.

The biggest thanks of them all goes to my family. My mom, who always pushed me, and played a pivotal role in getting me into philosophy in the first place. She was there to push me out the door to get to the exam that eventually granted me the admission to the Philosophy program at the University of Warsaw, even though I was close to giving up. Without you, your perpetual, even if often rough, love and support I would not be where I am now. I will be eternally grateful for everything you have done. Thank you mom, this dissertation is for you. Big thanks also goes to the

Italian part of my family. Laura, Marco and Paolo, you were always there for me, offering your help in any way you could.

And last but not least- Alex and Jan. I have never though that I could find someone that completes me in the way you do Alex. You have changed me in more ways I thought possible, you made me happy and whole, and I cannot imagine doing all this with anybody else but you. Through the PhD process, and a teething baby, you were always there to support me, being a partner, cracking jokes and making sure I have my coffee. I look forward to every step of the journey from here on, because I know there is no other person I would rather spend my life with. Jan, you tiny ray of sunshine, you gave my life new meaning. Both of you were my anchor when I needed one, and this dissertation is also for you.

Anna

Table of Content.

Road Map of the PhD.

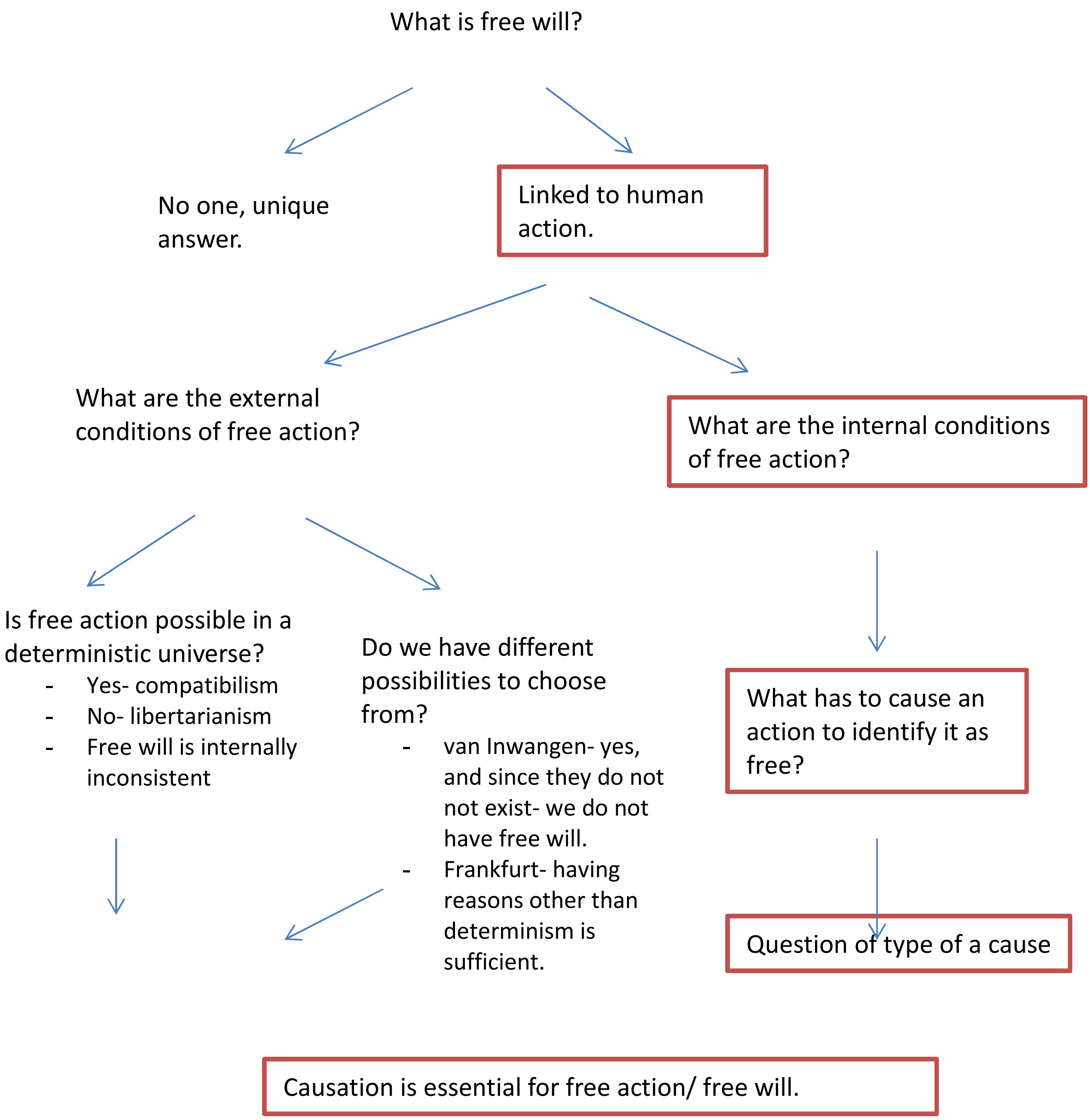
While the thought process that is behind any dissertation is very clear to the author, this is usually not necessarily the case for the readers. The text aims, of course, to guide the readers, step by step, through the thought process of the author, but, in case of dissertation especially, I was told visual aids are key. Therefore, before the discussion begins, I have mapped the most important steps that led me to writing this text. While, on their own, they might be simplistic, their aim is to guide the reader through the work of past four years, of which this text is a result. They are also aimed to be helpful, when, and that happens with every long text, the readers might feel lost, or find themselves thinking *How did I get here?*

The most important issues and questions are marked red, as they were central to the development of the main hypotheses.

Main goals of the dissertation:

- To show the connection between the problem of free will and the causal efficacy of the mental (mental causation as a necessary condition of free action).
- To present my own solution to the Causal Exclusion problem (to show that free will is possible).

First step:



What is free will?
No one, unique answer.
Linked to human action.
What are the external conditions of free action?
What are the internal conditions of free action?
Is free action possible in a deterministic universe?
- Yes- compatibilism
- No- libertarianism
- Free will is internally inconsistent
Do we have different possibilities to choose from?
- van Inwangen- yes, and since they do not not exist- we do not have free will.
- Frankfurt- having reasons other than determinism is sufficient.
What has to cause an action to identify it as free?
Question of type of a cause
Causation is essential for free action/ free will.

Second step:

What do we mean by causality when we talk about free will?

What makes a relationship causal?

What type of causes are we after?

Necessary

Law based approaches?
- co-occurences backed by laws
- Counterfactual dependences

What kind of link can be found that connects a cause and an effect:
- Processes
- Mechanisms
- Making a difference

Sufficient

Necessary and sufficient

INUS conditions type: a causal field of causes

Does not seem to determine what makes an action free

If it is not the type of causal connection, is it the type of a cause that makes an action free (or not)?

Third step:

What are the causes that result in voluntary, free action?

Neuroscience, e.g. Libet:
Action is voluntary and free if it is caused by a conscious decision, and not by an unconscious physical process

Belief/Desire model (standard approach), e.g. Davidson; the pair constitutes a reason for action.

Mele, Pacherie, voluntary action is caused by an intention; different potential classifications of those

Nanay: mental imagery. To decide (which is necessary for a voluntary action is to imagine oneself in outcomes of different situations_

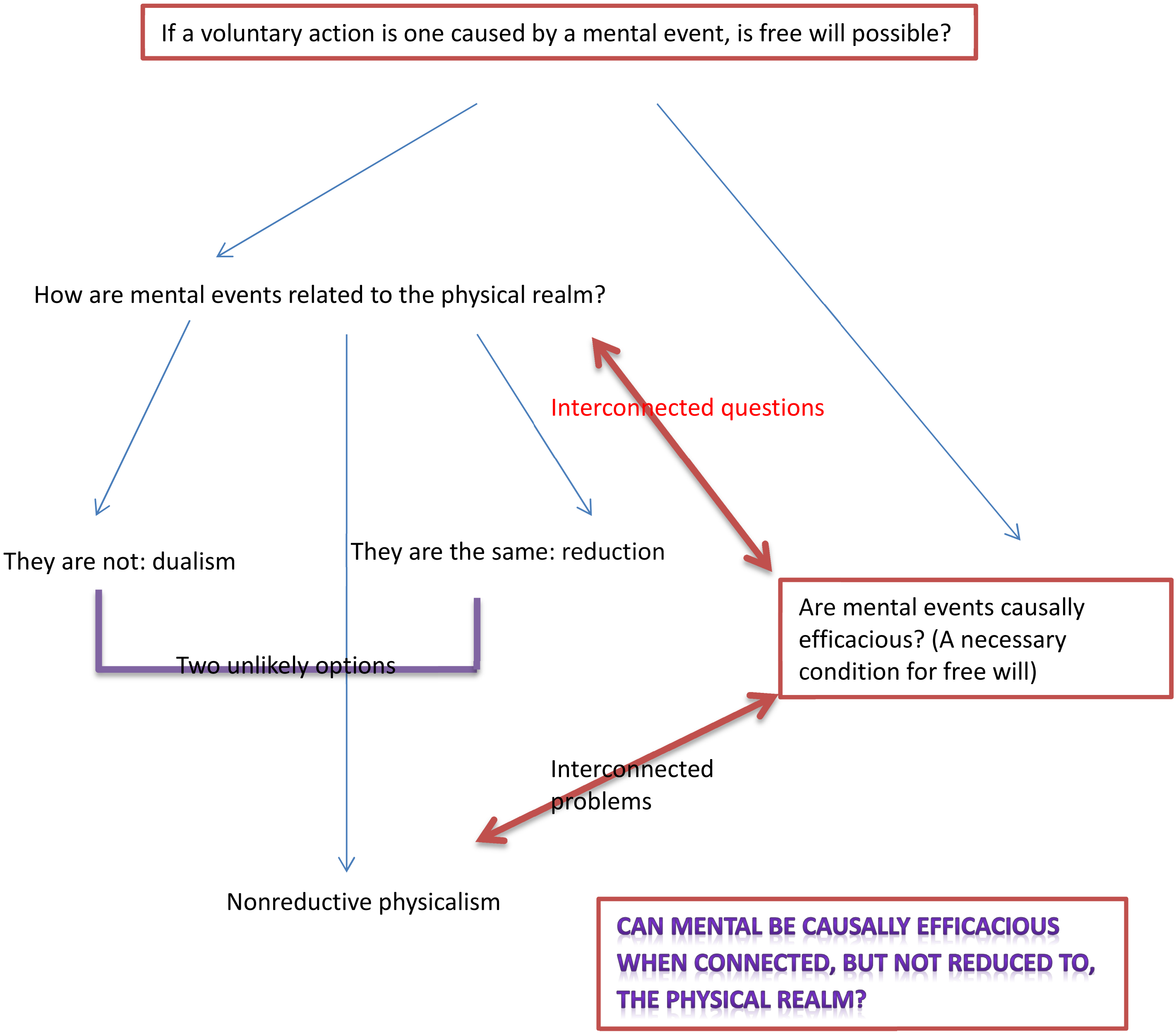
Do all intentions lead to voluntary actions?

Mele: yes

Pacherie: potentially no? There seems to be a qualitative difference between motor and distal intentions

One common thread: the mental character of the potential causes

Fourth step:


If a voluntary action is one caused by a mental event, is free will possible?
How are mental events related to the physical realm?
Interconnected questions
They are not: dualism
They are the same: reduction
Two unlikely options
Are mental events causally efficacious? (A necessary condition for free will)
Interconnected problems
Nonreductive physicalism
CAN MENTAL BE CAUSALLY EFFICACIOUS WHEN CONNECTED, BUT NOT REDUCED TO, THE PHYSICAL REALM?

Fifth step:

Can the mental be causally efficacious if it is connected to, but not reduced to the physical?

No, given the Causal Exclusion Argument.
What does this mean for the possibility of free will?

(1) An agent's action is free only if it is caused by the agent's mental states.
(2) Any effect that has a cause has a sufficient physical cause occurring at the same time (physicalism).
(3) The agent's mental states are not identical to any physical states, but supervene on underlying physical states.
(4) If an effect has a sufficient physical cause C, it does not have any cause C* distinct from C occurring at the same time (except overdetermination).

No action is free

Conclusions:

Therefore, for free will to be possible we have to:

Have a causally efficacious mental, which requires:

Solving the Causal Exclusion problem

Current most popular solutions:

1. Interventionism (e.g. List and Menzies, Woodward)
2. Causal compatibilism.

Both unsatisfactory

My solution

Pushing Mackie forward:
Can we talk about two causes?

Are our current accounts of causality able to handle the specific relationship between the mental and physical?

Chapter 1. Free will: Introduction to the problem.

There are undismissable philosophical questions- 'Do we have free will? is one of them- that require clear, well supported, soundly reasoned answers. We should not be bought off with allusive, impressionistic answers, however appealing or moving they may be. (Dennett, 2013, p.45)

In our daily existence it seems obvious to most of us that we are in charge of our own life and choices, to the point we rarely question it. We believe that we actively decide between various options, which way to go, which action to execute. We believe that there are different possibilities, different possible courses of action and it's up to us which one of those we will choose. We think that if we apply ourselves, we can improve our chances in life, our decisions have impact not only on our immediate but also long term future. When we look back, we see an array of events that we see as consequences of our choices, some of them we regret and we are happy with others. This view, that we live in a world where our choices have causal impact on our actions, could however be false.

A number of problems with the way we view ourselves as agents in the world can be contested on a few levels, and they are all, to some extent, covered by the different aspects of the problem of free will. The free will puzzle has been a centre of philosophical and theological discussions for centuries, admittedly the two often reaching different conclusions. The new interest for the old issues spiked with the introduction of the results of experiments, conducted mainly in neuroscience, which question the existence of free will altogether. However, the problem itself is not unambiguous, and encompasses an array of issues, all gathered under the "free will problem" umbrella, and it encompasses a number of issues linked to the questions about the nature of the universe, the causal source of our actions, the problem of moral responsibility. Answers to them all have the potential to change our perception of ourselves as agents, how we see other and how we operate in the social setting. Ultimately, they have potential to change how we see the societies and the law that governs them, making this challenge having potentially dire consequences.

It has been widely discussed that our common intuitions have been shaken by the results of the empirical tests. Is neuroscience going to solve this ages old question? I this dissertation, instead of settling on providing a conclusive answer, I will focus on both: the problems with the way experiments are designed, and their implicit assumptions regarding what would have to be the case for an agent to be free, and his actions voluntary. The general aim of this dissertation is to: (1) move the focus of philosophical discussions from the current one, where the possibility of free will is discussed in connection to the nature of the universe (deterministic or not) to the problem of internal conditions for free will by (2) arguing that causally efficacious mental is assumed in both the discussions in philosophy as well as in the way the experiments are designed, which (3) makes our notion of causation to be paramount, and (4) possibly forces us to look for solutions to the Causal Exclusion problem.

This work will begin by analysing what is usually meant by free will. In spite of the term being so broadly used, it comes with very strong intuitions, which might be a reason why it is rarely rigidly defined. But even relying on folk intuitions, different sub-problems come into foreground, including: control, self-determination, autonomy, choice, decision etc., which are not only not synonymous, but some of them have widely different meaning (e.g. decision seems to have more to do with self-determination than with autonomy etc.). Therefore, in order to meet my goals, I will begin by providing a background of the usually discussed problems, including the more classical issues (like free will's compatibility with determinism, and the problem of alternative possibilities), and the problems outlined in the more recent experimental developments. My aim here will be to identify what kind of conditions, both given explicitly, as well as seemingly assumed implicitly, have to be met, in order for free will to be a possibility. I will identify, and then clarify, notions essential to the debate, such as intention, reason and decision, arguing that all these events, are in fact mental, and mental causation lies at the core of the problem both for philosophers, as well as scientists.

In the following chapter (chapter 1) I will begin by clarifying what are the different understandings (section 1) and applications of the term free will, and I will discuss the, potential, differences between moral responsibility and free will (section 2). I will conclude the chapter by arguing that the discussion on free will, and its results lead to real live consequences, as, based on the studies conducted, the lack of belief in the notion can result in a change in behavior. Such results point to the importance of the debate, and the sensitivity of the topic.

What is all the "fuss" about? The general scope of the problem.

Intuitively, free will has something to do with our actions, and seems like a uniform issue, but to think of it this way would be a mistake. It is far from being a unified, homogenous problem, but rather is an umbrella term, covering a range of issues, from moral agency, human freedom and self- control, to the relation between mind and body, and others. Robert Kane identifies at least eight different issues that are connected to the discussion on free will:

1) moral agency and responsibility, dignity etc.

(2) the nature and limits of human freedom, autonomy, coercion and control in social and political theory

(3) compulsion, addiction, self-control, self-deception and weakness of will in philosophical psychology

(4) criminal liability, responsibility and punishment in legal theory

(5) the relation of mind to body, consciousness, the nature of action and personhood in the philosophy of mind and the cognitive and neuroscience

(6) the nature of rationality and rational choice in philosophy and social theory

(7) divine foreknowledge, predestination, evil

(8) general metaphysics issues about necessity and possibility, determinism, time and chance, quantum reality, laws of nature, causation and explanation in philosophy and sciences (Kane, 2002, p. 4).

The different issues, combined under this umbrella term, are vastly different from one another, and are usually dealt discussed not only in philosophy but also in, for instance, anthropology and theology. Most of the arguments in philosophy explicitly referring to the issue of free will are centred on the metaphysical problems, including the questions of necessity and possibility as well as determinism, laws and causation. Additionally some combine those issues with the questions about moral responsibility and moral agency, treating the term free will as synonymous with morally responsible.

Probably most widely discussed issue in connection to free will is its compatibility with the deterministic nature of the universe, where some (e.g. van Inwangen) question it is possible for free will to exist in the world where everything is determined by the antecedent state of the world in combination with the laws of nature. Clearly, determinism is here understood as causal determinism, a combination of all the notions clustered together in Kane's division above. I will treat all these discussions as concerning external conditions for free will.

Following this division further, the approach in science, and specifically in neuroscience, is to focus on free will "in action", and emphasizing the way actions or behaviours are caused. The authors here hypothesize that in order for an individual to be considered free, his actions cannot originate from some unspecified, unconscious brain activation but should be a result of some conscious decision. Each movement therefore has two potential causes: the physical, neuronal processes that are beyond agent's conscious cognition on one hand, and the conscious, probably mental, processes of intention and decision making The scientific approach to free will although, as I will hypothesize, operates under similar assumptions about the nature of choice as philosophers, has more in common with discussions in philosophy of action on action in general, than the ones explicitly discussing free will. I will treat this approach as examining internal conditions for free will.

Therefore, the external conditions for free will, will be concerned with the way the world has to be, in order for the possibility of free action to exist. The internal conditions, on the other hand, will focus on the inner states of the agent that would make free action possible, and here usually being caused in the 'right' way is at the centre of the debate.

Another angle in philosophical discussions on the topic include, besides the usual "arm chair" theories, empirical data obtained from the experiments on folk intuitions regarding free will. Although, providing different type of data, the focus is again mainly on the compatibility of free will with determinism. The upside of such approach is further clarification of how the notion is understood, and what might be some implicit assumption we all are operating under, offering a deeper understanding of the concepts.

Given those different focal points of the discussions, from internal to external conditions, it is not surprising that the rigid definition of free will, preferably one broadly accepted, is rarely given. Mele (2014a) identifies at least three different groups of definitions, each one imposing stricter conditions for the possibility of free will. In the first, most general group requires making rational choices that motivate the decision. It is enough that the individual makes rational decisions that he can justify. This broad definition is probably how most of non-philosophers think about free will. In the second group, besides the rational choice, also existence of alternative possibilities is required. The agent not only has to rationally choose something, also alternative options have to be available to him, which might be impossible given the deterministic nature of the universe. Therefore some form of indeterminism might be necessary. The third group links free will with having a soul or another supernatural aspect of life. Here all the dualistic approaches would fit. All three groups assume free action requires some form of control of the individual over his own behavior. All three are applicable to both the arguments trying to discover the nature of free will, as for the ones trying to investigate it in action, and all of them assume that free will involves some form of choice or decision.

There are also those (e.g. Pereboom, 2001) who emphasize the control aspect of free will, we are free when we are able to control our actions and our behavior, and then if factors beyond our control are the cause of our behavior (e.g. the antecedent state of the universe combined with the laws of nature), and we are not in control of those, we are not free. The notion of control, although important, seems to be based on a slightly different premise than the one required for the rational explanation of the action, as it is perfectly plausible that the agent might have reasons to do something, while at the same time, the action was beyond his control.

In this dissertation I will focus mainly on choice, although, as we will see later, the condition of control can present seriously issues to choosing and action initiation processes. As we will see in the discussions further, I will discuss both the external as well as internal conditions of existence for free will, however I will argue that the latter are more pressing. Throughout I will operate with the idea that, in order to consider an action free, certain internal conditions have to be met.

Free will or moral responsibility?

Often in literature free will is discussed with connection to moral responsibility, and, in some cases, the notions seem to be used almost synonymously. However, as we have seen in the previous part, different conditions can be identified as necessary (although possibly not sufficient) for free will to be possible, which might complicate the assumed synonymous nature of the terms. When can a person be held morally responsible? Manual Vargas defines moral responsibility though the justification of our moral practices:

The work of the concept of moral responsibility has to do with the marking a set of inferences about differential moral praiseworthiness and blameworthiness. People who *knowingly and intentionally do the wrong thing derive condemnation. People who do the right thing deserve approval. Keeping track of when people deserve praise and blame is the work of the concept of moral responsibility. It is, in some sense, why we rightly have a concept of moral responsibility* (Vargas, 2010, p.61)

In this sense, it seems that moral responsibility has more to do with the free will in action, or free will as it is exercised, rather than with the concept of what free will actually is. The passage quoted above, suggests that we can be morally responsible for our action based on how, or by what, said action was caused. Moreover, the praiseworthiness and blameworthiness are only determinable after the action has already occurred, and the agent can identify his intentions or reasons, that guided him in choosing this specific action. Therefore, although the universe has to be such that different actions are possible, what in the end might be the core is the "right" cause, reason or intention.

Three main approaches to moral responsibility with its connection to free will can be identified as (Timpe, 2013): (1) the reactive attitudes, (2) the ledger view, and (3) the accountability view.

The reactive attitudes approach was developed by Peter Strawson (1962), who defined moral responsibility through its connection to social practices. If we think that an individual is morally responsible we tend to have specific attitudes towards them, namely reactive attitudes. They are reactions we display toward others, depending if we think their behavior is a result of a good will or to the contrary, ill intentions.

The ledger view was defined by Zimmerman (1988) and ledgers are metaphorical "marks" a person gets whenever they behave in a specific (praise or blameworthy) way. If the behavior is deemed to be positive the metaphorical record of the person gets upgraded, and a negative behavior leaves a negative mark. Although such record is nothing more than metaphor, for Zimmerman, it captures the core of our understanding of moral responsibility.

The accountability view (e.g. Oshana, 1997) is based on the assumption that moral responsibility is the ability of the agent to provide an explanation for their behavior.

These views of moral responsibility neither aim to prove that free will exists, nor give a comprehensive definition of the notion that would be, when applied to the moral responsibility problem, or independently of it. Rather they focus on what behaviors are generally accepted as good and bad, and what are the conditions under which we have the tendency to praise or blame other individuals.

In this dissertation, although I do not think the two notions are necessarily synonymous, I will treat them as such, and whenever moral responsibility is discussed I will treat the argument as one concerning free will. My focus seems to fall in line with the way the issue is discussed in the positions mentioned here, where what matters is how the action was caused. Therefore, I assume that both for the accounts of moral responsibility, and, as I will argue here, free will, the central issues is the specific way the action was caused, making them, for the intents and purposes of this work, synonymous. If we look at the positions discussed here, both the praiseworthy as well as blameworthy actions are distinguished not based on how they are caused, as they are all caused by, what I argue is a mental event. They are divided based on the content of said mental event, but they are all caused in the same manner. Therefore, since here my main focus is on *how* the action has to be caused in order to be identified as free, it makes sense to treat the two concepts, free will and moral responsibility, as synonymous.

Is it important that we believe in free will?

What difference does believing, or disbelieving in free will make? The belief that we are free forms a basis of our intuitive understanding of humans and agency, but what are the consequences when this view is shaken? What would be the consequences if individuals would stop believing in free will, regardless of the truth or falsity of its existence? If the conclusions of the neuroscientific experiments would be universally accepted, would it have an impact not only on our laws and society, but on the behavior of the individual himself? Since those kinds of statements are already making their way into the popular magazines and newspapers (e.g. an article from TIME magazine *Can Neuroscience Debunk Free Will*, but there is many more), some decided to investigate if our behavior changes when we stop believing in free choices.

Daniel Dennett in one of his articles (2014) illustrates the problem with a story of a neuroscientist who, after performing a brain surgery on a patient, informed him that the chip, placed in his brain to control his obsessive- compulsive disorder, besides performing what it was designed to do also controls all his choices and decisions. In reality, there is no such device, and the chip placed in the brain had only one, standard function: controlling the obsessions, as was agreed upon before the surgery. What happened next? Dennett speculates:

It worked; the poor fellow went out into the world convinced that that he was not a responsible agent, but rather a mere puppet, and his behavior began to show it: He became irresponsible, aggressive, negligent, indulging his words whims until he got caught and put on trial. (Dennett, 2014, p.77)

The belief that something else, rather than him controls his behavior, in Dennett's scenario, led the patient to change his behavior and ultimately lead him to crime. During his defense the men tried to explain that all his action was caused by the implant in his head, but the neuroscientist when questioned admitted that it was a prank she did not think the men would believe.

At a first glance the story sounds a bit like a science fiction movie, and true, we do not have the technology advanced enough to trick somebody to believe we could in fact control their behavior through a chip implanted in their head. But Dennett example expresses certain intuitions: that belief in free will makes a difference to a person's behavior, which were shared by scientists, and led Vohs and Schooler (2008) to construct and conduct a series of experiments investigating them.

In their study Vohs and Schooler aimed to discover potential links between believing in free will and cheating. On their approach free will was presented as incompatible with determinism, therefore the scenarios in which determinism was true and everything was a consequence of environmental and genetic factors, were treated as unfree. The study aimed to further investigate the link between the sense of responsibility, and the behavior, where usually the former is believed to heavily influence the latter (e.g. Harmon-Jones and Mills, 1999; Ajzen, 20012; Bandura, 1982).

Two experiments were conducted. In the first one two conditions were tested:

(1) Anti-free will, where participants were reading a passage from a book by Francis Crick, where it was clearly stated that free will is illusory and a creation of our own minds
(2) Control condition, where participants were reading passages from the same book, but from a chapter about consciousness instead of free will. In those chapters, there was no mention about free will.

First, before the experimental part, the participants were asked to complete the Free Will and Determinism (FWD) Scale and the Positive Negative Affectivity Schedule (PANAS) (both are questionnaires regarding personal attitudes towards the belief in free will in this case, usually the answers are given on the 5 step scale, Voohs and Schooller for the FWD used the study by Paulhus and Margesson, 1994 and for PANAS: Watson et al. 1988, more detailed description to be found in Annex 1.). After completing the questionnaire, the participants were asked to fill out a computer based test, which, they were told, had a programming glitch causing the correct answer to appear on the screen while they were answering the question. They were also told that it was possible to stop the glitch by pressing on the space bar, after the problem to be solved appeared on the screen. They were additionally told that the investigator would only see the answer, and would not be able to tell if the space bar was pressed or not, while in reality both the answer and the number of space bar presses were recorded.

Results of the first experiment showed conclusively that the participants who read the chapter falsifying the existence of free will, cheated more frequently than the ones from the control group.

In the second experiment, the participants were divided into five groups, corresponding to five conditions. From those five, in three cheating was possible. Similarly to the previous experiment, also here the participants were asked to read a text aiming at either increasing or decreasing their belief in free will. In the free will-condition the text included sentences such as:

I am able to override the genetic and environmental factors that sometimes influence my behavior. (Vohs, Schooler, 2008, p.51)

In the deterministic conditions participants read sentences such as:

Ultimately, we are biological computers- designed by evolution, build through genetics and programmed by the environment. (Vohs, Schooler, 2008, p.51)

While in the neutral conditions the statements were similar to:

Sugar cane and sugar beets are grown in 112 countries (Vohs, Schooler, 2008, p.51),

which clearly had nothing to do with free will.

Afterward the subjects were again asked to first complete the two questionnaires, and then a computer based test. They were then told the experimenter will give them feedback on the results of the test, and they will be paid one dollar for every correct solution to the problem from the test. After giving the subjects the instructions, the experimenter stated that he forgot about an important meeting and had to leave, therefore the participants should check the test themselves, score their own problems, pay themselves for every good answer and then destroy the sheet, since the researcher did not have the right to keep them.

After the results of the test were analyzed, it was established that the participants that took part in the determinism condition, paid themselves more money than the ones participating in all other conditions. It was again concluded that lack of belief in free will leads to cheating behavior.

In other studies (Leotti et al., 2010) control over our actions and the ability to influence our environment is described as essential for our survival, both for people as well as for animals. Ability to change our environment is understood as a form of control over it, and could be seen as not necessarily synonymous with free will.

Based on experiments above, we can see a clear link between the individual's behavior and they belief in the existence of free will. Such belief was also found to have links to an increased helpfulness and reduced aggression, while decreasing that belief led to more selfish and automatic responses (Baumeister et al., 2009).

The studies discussed here seem to all agree that our belief in free will, regardless of its actual existence, has real-life consequences. The decreased belief in our agency has very tangible influence on our actions, and our compliance to follow the societal norms. These results show that, besides being scientifically interesting, the discussion has consequences to the way our society functions.

Conclusions.

This short chapter showed that free will, although being at the center of a broad and versatile debate, that includes researchers from a variety of domains, lacks clear and rigid definitions. The approaches can be divided into several different groups, based on the conditions that are seen as necessary for the existence of free will.

I have discussed the difference between two notions: free will and moral responsibility, and argued that for the intents and purposes of this dissertation I will treat the two notions as synonymous and interchangeable (although they not necessarily are).

Moreover studies were shown, that hypothesize a link between the decreased belief in free will and an agent's behavior. The results of philosophical discussions, and the neuroscientific and psychological experiments, are then taken to have real life consequences, emphasizing the importance of an in-depth and interdisciplinary debate.

Short summary of my position as presented in this chapter:

1. My approach to free will fits into the broadest definition, of the three presented by Mele, for a very specific reason: in this dissertation I want to argue that the problem of the internal conditions of free will is more pressing than the external conditions. Therefore I am, for now, remaining agnostic on the matter of the necessity of indeterminism.
2. I will treat moral responsibility and free will as synonymous, as far as the focus is put on the way the action was caused. I will not analyze further possible problems

linked to the synonymous, or not synonymous, nature of the connection between free will and moral responsibility.

In the next chapter, we will see the main lines of discussion found in philosophy, focused, as I discussed previously, on the external conditions that need to be met for free will to be a possibility.

Chapter 2. Main axis of discussion. Free will as philosophical problem.

The problem of free will has been present in philosophical discussions for centuries, and currently most of the debates are focused on what I called the external conditions for free will. Such conditions are most accurately described as the compatibility of free will with the potentially deterministic nature of the universe, and form the first axis of the discussions. For some the exact nature of the universe remains to be determined, with both determinism and indeterminism remaining in the realm of possibilities, while others (e.g. libertarians) argue that free will is only possible with one of them being the case.

The second axis is centred on what has been called the principle of alternate possibilities, where the main issue is the existence of actual possibilities for an agent to choose from, as a condition of possibility of free will. The two problems, that of determinism, and that of different possibilities, although not the same, are concerned with similar issues. In case the universe is in fact deterministic, it seems to exclude the existence of different options. However as we will see some (e.g. Frankfurt) question if those possibilities are in fact necessary, or is it sufficient for the agent to have reasons for executing the action in question.

Frankfurt's argument, aiming to provide a solution to the problem of alternative possibilities, leads us to the third axis of the discussion, which connects the problem of free will with the causal origin of our actions. Here the actions to be free, they have to be caused in the "right way", by reasons or intentions. Conditions of existence of free will would be defined here through free action, which in turn would be a result of a specific set of causes. The questions about what causes our, not necessarily free, actions are more prevalent in philosophy of action, while in connection to the former notion are to be found mostly in neuroscientific experiments. There the action can be identified as free, when it was caused by a conscious decision, intention or reason and not by the physical, and unconscious brain activation. Although this approach is centred on the internal conditions of free will, it is intertwined with the question of determinism, with some (e.g. Nahmias et al. 2007) arguing that the truth of determinism means that the mental is epiphenomenal.

The three distinct axis are interconnected, although it is usually the first two (compatibility with determinism and alternate possibilities) that are covered in detailed, and are discussed together. Often in discussions on free will, their starting point is the assumption that given certain conditions, the context of action, the agent's desires etc., we would have to be able to do something different than what we have actually done. That if we have chosen to go by bike to the train station and not with the car we would be more fit, or that if we would decide to marry a different person than the one we have actually married we would not have been as happy as we currently are. Those decisions do not have to be major life choices, they can be mundane small actions. If I would have decided to put the water into the fridge not I could enjoy a cold drink, but since I didn't all I can use to satisfy my thirst is a glass of tepid water. Then the way the world would have to be in order for those assumptions to be true, is analysed.

Being able to do something different that we had in fact do is a belief that most of us seem to share. Yet, as we will see below, it can be very difficult to argue for, on philosophical ground. Given that the world is usually considered to be deterministic, how is it possible for us to do something other than what we actually did? Is there space for choice when determinism is king? Another issue is if those alternative options, which we identify as possible futures, or possible different decisions, have to actually exist. This commonsense conception about the openness of our future, is challenged by van Inwangen's argument, where the deterministic world leaves no room for possibilities. However, as we will see blow, Frankfurt thinks that for free will to be preserved, an illusion of choice might be enough.

Between the two frameworks of the discussion on free will, its compatibility with determinism and alternate possibilities, very close links can be drawn, while the link between the two and the third axis of discussion seem to be a bit looser. The latter rather than analysing the external conditions for free will, deals with the causal origin of our decisions, investigating what kind of things have to happen before the action and where does the action originates, which I have described as the internal conditions. It allows to, not only, specify under which conditions free will is possible, but also what free will might be. Neuroscientists usually assume that for an action to be classified as free, it has to causally originate from a conscious decision rather than from an unconscious physical process, suggesting that the two differ in some way. Setting aside its exact ontological status, it suggests that the conscious decision is mental, linking the problem of free will with the issue of mental causation. Between the first and the third axis of discussion, the convergent points are, among others, the notion of causation. Both in case of arguments concerning the nature of the universe, as well as those focusing on the necessity of being caused in the 'right' way, the focal point is the notion of causation. In the former case, it is the possibility of non-determined way in which our actions can be caused, while for the latter it is the question what are the causes of, presumably, free actions. For the second and the third axis the converging question seems to be what causes the free action: the deterministic, no alternative processes in the former, and again, the type of the cause necessary, for the latter.

The next sections of this chapter will focus mainly on the first two discussions: (1) compatibility with determinism, and (2) alternate possibilities. In section 1 I will analyse what is usually meant by the term determinism in the discussions on free will, and what that would mean to our understanding of our actions if the world would in fact be this way. In the further sections I will discuss the three most popular positions on the matter with: (1) compatibilism, (2) libertarianism, and (3) hard determinism. I will point out the advantages as well as shortcomings of all three. It is important to emphasize that, although I think that the problem is serious, here I choose to remain agnostic as to the problem of free will's compatibility with determinism.

I have mentioned in the introduction that the main focus of this dissertation is, at least, twofold: (1) show that causality, predictability and determinism, while usually discussed together, come with independent problems, and (2) that the way things are caused matters for free will (I call it the internal conditions for free will). Those have to be met regardless of the nature of the universe, and therefore constitute a problem for all, regardless if they think the world is deterministic or indeterministic. I will also discuss some of the results obtained by experimental philosophers concerning the compatibility of free will with determinism. Next, in section 2.2., I will proceed to discuss the principle of alternate possibilities, and the arguments surrounding it, which will conclude this chapter.

The threat of causal determinism. Is free will possible if the universe already "decided" how your life will look like?

We are used to operating within a reality where we make choices, and we are responsible for the actions that we conduct. Form simple movement of the hand to complex life-changing decisions, we believe that the ultimate outcome is dependent on us as agents. Potentially if we accept that universe is deterministic and governed by a set of laws of nature, this view might not be supportable. The type determinism discussed in connection to free will debate, otherwise known as causal determinism[1], is a major obstacle for both, showing how choice would be possible, as well as the existence of different possibilities to choose from.

Historically the discussion on the problem if our behaviour is 'up to us', or if the origin of our action lies elsewhere, can be traced back to Aristotle. Determinism, as probably most of the notions out there, can be understood in a number of ways, with four of them seemingly used most commonly: (1) logical determinism, which stipulates that the truth values of all the propositions, including the future ones, are fixed and either true or false; (2) theological determinism accepts the existence of an all-knowing God that has the knowledge about all the past, present and future events; (3) temporal determinism, as formulated by T.S. Elliot treats time as another dimensions where both "sides" of time are the same[2]; and (4) causal determinism, where all the past events combined with the laws of nature entail all the future events. This type of determinism is usually considered problematic for what we deem important for free will (e.g. choice between different options, being able to determine own future etc.).

There is a number of ways in which causal determinism can be defined, where the broadest ones formulate it, more or less, as follows:

[…] given the physical laws and state of the universe at some time, the state of the universe at all the other times is uniquely specified (Roskies, 2014, p.104)

---

[1] In this dissertation I will understand the term determinism when discussed in the framework of the free will problem, as causal determinism, and I will use the two interchangeably.
[2] More on temporal determinism in Campbell et al. 2004

This, compact, definition not only specifies which characteristics of determinism are problematic, but also suggests what that entails: the unique specification of the events. Therefore the combination of the physical laws, with the knowledge of its antecedent conditions, at a chosen moment give rise to only one, specific, situation. The universe being specified at all other times, should be understood as determination of not only the past, but also all the future events, however it usually is taken not to entail the predictability of the events.

This broad definition, focusing on the physical laws and the state of the universe, will be sufficient to understand most of the arguments in the free will debate. However, similarly as with the concept of free will itself, the definitions used often reflect other assumptions, and Robert Kane argues that an event can be considered determined:

[…] just in case there are conditions (e.g. the decrees of fate, the foreordaining acts of God, antecedent physical causes plus laws of nature) whose joint occurrence is (logically) sufficient for the occurrence of the event: It must be the case that if these determining conditions jointly obtain, the determined event occurs. (Kane, 2011, p.4)

Kane's formulation extends the threat of determinism to include also threats coming from God or fate, rather than only antecedent physical conditions, as is usually assumed in the modern discussions. The occurrence of the event is then inevitable, given the conditions described above, leading to one of the main (seeming) consequences of determinism, namely predictability.

Probably the most popular formulation of causal determinism comes from Laplace, although in his case, the deterministic nature of the universe entailed its perfect predictability. On causality and causal chains he wrote:

Present events are connected with preceding ones by a tie based upon the evident principle that a thing cannot occur without a cause which produces it. This axiom, known by the name of the principle of sufficient reason, extends even to actions which are considered indifferent; the freest will is unable without a determinative motive to give them birth (Laplace, 1902, p.3)

All events, for Laplace, are therefore connected with the previous events in the causal chains, and the present state of the universe is considered to be:

[…] the effect of its anterior state and as the cause of the one which is to follow. (Laplace, 1902, p.4),

further emphasizing the interconnected nature of causal chains, which are wide spread throughout the universe. Those interconnections however have one, problematic, consequence, which is seen as a threat to the possibility of free will, namely the predictability of the universe. In the famous quote Laplace writes:

Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it- an intelligence sufficiently vast to submit these date to analysis- it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. (Laplace, 1902, p.4)

Therefore, given a vast enough knowledge and comprehension of the universe, there is a theoretical probability a super being could predict an outcome of every event, based on the antecedent state of the universe, combined with the laws the universe follows. If the future is perfectly predictable from the combination of antecedent conditions and the laws of the universe, no space seems to be left for free will, which seems to require at least a moderate level of variety in the future events. Laplace however did not claim that we would ever possess the skills, or the intellect, to be able to predict the outcome of every situation. Rather a perfect being, close go God, capable of assessing all the variables at any given moment, and all the events in the future, would be able to also predict all the events in the future. Laplacean demon has been hunting philosophers working on free will ever since.

If everything is a consequence of the previous states of the universe and the laws of nature, no “wiggle room” is left for free will, as everything has already been fixed around the time of the Big Bang. The future is not open, but has to follow a pre-determined path, individuals cannot act differently than the way they acted, since also their behaviour is a result of the previous states of the universe and the laws of nature. At the core of the free will lies the idea of an agent, that is able to choose how to act and choose between different possible scenarios, the vision of the future that is completely fixed is not compatible with that picture. The argument can be easily summed up in three principles:

(1) *If determinism is true, then we are never able to do otherwise.*
(2) *If we are never able to do otherwise, we have no free will.*
(3) *Therefore, if determinism is true, we have no free will.* (Vihvelin, 2013, p.2)

Kane (2011) on the other hand argues that if any form of determinism would be true it would pose a threat to either one of the conditions for free will, where the action would no longer be:

(a) *"Up to us" what we choose from an array of alternative possibilities, since only one alternative would be possible* (Kane, 2011, p.5),
and that:
(b) *The origin or the source of our choices and actions would not ultimately be 'in us' but in conditions, such as the decrees of fate, the foreordaining acts of God, or antecedent causes and laws, over which we have no control.* (Kane, 2011, .p.5)

Therefore the causal origin of our actions is not 'up to us', but to the antecedent conditions combined with the laws of nature. Defined this way, determinism does not necessarily entail, or enforces talk about predictability, but, as we will see further, it often accompanies it. This often comes from the folk, commonsense understanding of the consequences of determinism. However, while it show that the two main axis of discussion are usually connected, it helps to show that causality has its own standing. Our actions have to, presumably, be caused in a specific way, in order for them to be free. This condition holds for an action generated in either deterministic, or indeterminisitic manner. In case the world is deterministic, our intentions, if causal, can cause our action in a deterministic manner: the previous state of the universe (including the intention itself and the causes of the intention) and the laws of nature (although here the difficulty might be identifying what the laws would be in connection to the mental) would determine what is the outcome caused by the specific intention. If that supports free will is a different question, here the main issue is if intentions or reasons can possibly be causal at all, and if the mental is not epiphenomenal, existing and not reduced to the physical, but also not causing anything.

The discussions on the challenge from determinism do not necessarily claim that either determinism, or indeterminism, are in fact the case, although some proponents of libertarian positions, state that given the advances of quantum physics we can argue that the world is in fact indeterministic. The question about the nature of the universe constitutes a purely empirical, and to date to some extent unanswered, question, but the answer to it has metaphysical consequences, discussed by philosophers. On the other hand, neuroscientists (e.g. Haggard, 2011), have the tendency to argue that the world, and neuroscience are deterministic, and therefore any account of free will, will have to be compatible with it. Others, like Gazzaniga, argue that although it is not set in stone, it makes sense for neuroscientists to make an assumption of determinism, and proceed with theories under such assumption. Gazzaniga argues that:

[…] it is widely believed that determinism holds at the level of neurons, even if not at the level of atoms. Neuroscientists believe that Newtonian- type deterministic laws reliably predict how the brain functions and produces mind. One could say this is an attitude of neuroscientists, a stance. In fact, it is way too early for neuroscientists to claim that they actually know how neurons produce mind. Still, it seems to make a lot of sense to believe that someday neuroscientists will understand how neurons interact to produce mind. (Gazzaniga, 2014, p.68)

The hypothesis that neuroscience will solve the problem of free will, and that they will provide a framework which explains how neurons produce mind is widely spread in neuroscience (besides Gazzaniga, also Tse, 2013). However the two issues are not synonymous, which it not always recognized. The reasons for it, including the very broadly accepted, implicit, dualistic view, will be discussed in detail further in this dissertation.

Others (e.g. Libet, 2004) treat both determinism as well as indeterminism as still unproven, metaphysical theories.

The current positions regarding the compatibility of free will with the deterministic nature of the universe can be divided into three main groups:

(1) Compatibilism, whose proponents argue that even if determinism is the case, free will is still a possibility.

(2) Incompatibilism, which claims that free will requires some form of indeterminism. Here the position can be further divided, as the (a) incompatibilists do not take a definite stance on the nature of the universe, while (b) libertarians state that free will requires indeterminism, and universe (most likely) is indeterministic. In this dissertation, I will use both terms interchangeably and will emphasize if the discussed position clearly distinguishes between the two.
(3) Hard determinism where the universe is deterministic and we do not have free will. Hard determinism is also an incompatibilist position, however in the text I will refer to it as hard determinism.

Compatibilism.

Although determinism remains only a possibility, some, most famously Dennett, argue that even though the world might be deterministic, there is still space for free will. This free will however might not look like what we expect it to. In response to arguments made by Wegner[3] (2002), refuting the causal power of conscious will, Dennett argues

If you are one of those who think that free will is only really free will if it springs from an immaterial soul that hovers happily in your brain, shooting arrows of decision into your motor cortex, then, given what you mean by free will, my view is that there is no free will at all. If, on the other hand, you think free will might be morally important without being supernatural then my view is that free will is indeed real, but just not quite what you probably thought it was. (Dennett, 2003, p.222)

Two separate issues can be identified here, firstly Dennett seems to reject the idea that free will requires dualism, and that we need a soul in order to be free. This assumption could be compatible with determinism, however is probably usually taken to work with indeterminism. Here I assume that Dennett, by rejecting dualism, shows that indeterminism is unnecessary, mainly because from the discussions it appears that the laws of nature that are being discussed are mostly taken to be about the physical realm. Dualism would then seem to stand in contradiction to those, and through that, with determinism. Important to note here, a full analysis of the concept of laws of nature that is used in the discussion would

[3] The argument by Wegner will be discussed with more detail in chapter 2.

require time and space unattainable in the span of one dissertation. Most of the time, when I refer to laws of nature in this text I mean strict, exceptionless generalizations, based on regularities, which, if at all, are probably discovered in the realm of physics. I am aware that this understanding has various issues, but it’s the broadest and probably the most applicable for the problems discussed here. I will also sometimes discuss ceteris paribus laws, but if I do so, it will specified in the text.

Second issue is the nature of free will and what it requires. Dennett does not specifically define free will, but he argues that it requires nothing supernatural, presumably assuming that if free will exists, it has to be compatible with how the world is. Therefore, if the world is deterministic, free will should be compatible with it.

In discussing the dangers of determinism, Taylor and Dennett (2011) argue that most of the proponents of incompatibilist approaches confuse two theses concerning possibility and causality (much like in the quote by Robert Kane I presented above). They claim that, on libertarian accounts, if determinism is true, then firstly no one can ever truly utter the sentence ‘I could have done otherwise’ and moreover no one can neither be blamed or praised for an action, as they were all predetermined. Those requirements, as the authors argue, can also be met if the world is deterministic, and the do not require any form of quantum indeterminacy. If they hold or not depends on how we define possibility and causality.

In our daily usage of the notion of possibility, even if the world is deterministic, our beliefs about the possibility of certain events do not change. The threat that determinism presents to free will relies on, not only, what we understand as possibility, but also what we consider to be antecedently sufficient conditions. Are they only physical micro states, or rather a broader context of psychological and cultural norms? If the latter, what is the relation between them (the physical realm and the psychological and normative realms), are they reducible to one another or a different relationship holds? Moreover, depending on what we are looking for, the microscopic causation might turn out to be uninformative, and the question ‘could it happen differently’ seems to hold little meaning to it. The only possibilities that do not exist are bound to that type of consideration based on looking on a micro-scale of distinct events. Even then, broader possibilities can nonetheless exist and adding indeterminism to the world would change nothing in this worldview, since we usually talk about indeterminism on the “lower level”. Similar path, where determinism is deemed irrelevant to the problem of moral responsibility was taken by P.F. Strawson, whose position was briefly mentioned above in connection to the relationship between free will and moral responsibility. On his account it’s the reactive attitudes that matter, and it does not matter if determinism is true, there are other reasons to hold subject’s accountable and responsible.

Another problem Dennett has with the deterministic threat to free will is the notion of determinism itself, which is in itself controversial. Thinking about determinism the way Lapleace has, is, according to Dennett, a mistake. He questions (Dennett, 1984) if it is even possible to claim that we can put ourselves in the place of Laplacean demon, which has the perfect knowledge of all the states of the universe at any given time, and if determinism can even be understood in that manner. Ultimately, even if there was such creature, we are not it, and therefore all the states of the universe and all the laws of nature are simply not accessible to us.

Dennett (2003) also reject the libertarian assumption that determination is synonymous with causation. He argues that determinism is not necessarily about causation but rather about sufficiency, and is compatible with, for instance, events that do not have any causes at all. He states:

To say that if determinism is true, your future is fixed is to say nothing interesting. To say that if determinism is true your nature is fixed is to say something false. Our natures aren’t fixed because we have evolved to be entities designed to change their natures in response to interactions with the rest of the world. It is confusion between having a fixed nature and having a fixed future that mismotivates the anguish over determinism. (Dennett, 2003, p.93)

From his perspective, we can only see the future from the position of the agent’s and not from the position of a God, or Lapeacean demon. Agent’s perspectives change all the time, and we have to meet those changes, and for Dennett:

The distinction between being a thing with an open future and being a thing with a closed future is strictly independent of determinism. In general, there is no paradox in the observation that certain phenomena are determined to be changeable, chaotic, and *unpredictable, an obvious and important fact that philosophers have curiously ignored.* (Dennett, 2003, p.90)

What seems to be guiding Dennett, and possibly other compatibilists is the conviction that, first of all, we have always been living in the same circumstances, physically speaking. Since those are the circumstances we have developed notions such as alternative possibilities, or free will, they should be taken seriously and not dismissed based on the conditions that have always been there. We have always had brain, the neurons were always firing, we are physical creatures, and if we have free will, it has to be compatible with the way we are made. Moreover, the answers we are looking for seem to be dependent on the context. This, as we will see further on in this chapter, was very important in Frankfurt's understanding of free will, as, although our action surely has a micro-physical explanation, it might not be very useful for us, in this specific context. Moral responsibility is a good example, as when we ask for reasons the agent had for a specific action, citing micro-physical explanation is not particularly helpful, although it is true. Distinguishing between the causal explanation, understood as a form of causal integration that happens between the agent's psychology, and the action itself is another way to avoid the challenge of determinism. These types of accounts (e.g. Frankfurt, Fisher) take our psychological reasons, and reactions to our actions to be central for moral responsibility. It does not matter (as in case of Frankfurt, discussed elsewhere in this chapter), if determinism is the case, or if our reasons were physically determined. What matters are our intentions, which can be used to determine when a person can be held responsible for their actions.

The last point can be interpreted to connect the discussions on determinism, and causality. Regardless of the true nature of the universe, even if it is in fact deterministic, for an action to be free, is seems, it has to be caused in a very specific way.

Libertarianism and the incompatibility of free will with determinism.

Probably the strongest argument against the compatibility of free will and determinism, the Consequence Argument, was presented by van Inwagen (1983), and reads:

If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (van Inwagen, 2013, p.117; 1983, p.5)

The expression used by van Inwagen is confusing, and again can be seen to show that the problem of causation is, to some extent, independent from the determinism/indeterminism debate. It seems up to us, because this is the part of the expression I am discussing here, can be read in a number of ways, including that an agent is a separate substance doing things, or that our actions have to be caused in a specific way to be free. The second interpretation lies at the core of this dissertation.

Different formulations of this problem were also presented by Carl Ginet (1966) and David Wiggins (1973).

The argument focuses of what determinism really entails, and van Inwangen argues that given the deterministic assumption, our future is not open, but was determined before our birth. This seems to stand in direct opposition to what we usually consider essential for free will, as it turns out that nothing is in fact up to us, all of our choices have been already determined before we were born and everything is based on the combination of conditions and laws of nature, which we have no influence on (this depends on how we think about laws of nature). In other words, the fact that I am right now sitting on my chair, writing my dissertation, at the time t_1 is a consequence of some previous state of the universe, at time t_0 and the laws of nature. Therefore, it is not up to me that I sit here writing, since, if I would want to change my current position (and be somewhere on a party), I would have to change the previous state of the universe and the laws, neither of which I can do. The previous state of the universe lies in the past so I have no influence on it, and I probably have just as little impact on the laws of the universe, as I presume anybody else has. Therefore, if the universe is in fact deterministic, I never had any say in how my evening today would look like, it was all set previously at a time most likely unknown by me, presumably the big bang.

As we have seen in the previous subchapter, compatibilists tend to reject this argument offering an interpretation of van Inwagen's argument in such a manner that the dire consequences are avoided, e.g. discussed here in different parts of this chapter Dennett and Lewis. Libertarians on the other hand seem largely convinced by said argument, and as a consequence, wanting to preserve free will, they deny the truth of determinism. Unlike incompatibilists, who only argue that free will is incompatible with determinism, but don't necessarily state that the universe is in fact indeterministic, libertarians combine the rejection of determinism with simultaneous belief in free will, meaning that, at the core of their arguments is the assumption that the universe is in fact indeterministic. Therefore for libertarians, it is not the case that the state of the universe at a given time, combined with the laws of nature, provides us with the unique outcome that is somehow unavoidable, and many of them argue that quantum physics provides a good reason to believe that the universe is in fact indeterministic. In such a scenario the indeterminacy at the lower, presumably micro-physical, level leads to freedom at the higher, presumably cognitive, one. This argument, as we will see further here, might be problematic, as indeterminacy is connected to randomness rather than freedom, and the two are not interchangeable, meaning that lower level indeterminism might bring us chaos rather than freedom.

Libertarian positions can be divided into two broad categories: (1) Agent-causal libertarianism: advocated by, among others, Timothy O'Connor (e.g. 1995, 2011), where it is not events but agents that cause actions, as, to some extent at least, independent substances; (2) Event-causal libertarianism, where only states or events can cause a behavior, and the production of the action relies on indeterminism being true. This type of libertarianism can be found in the work of e.g. Ginet and Kane.

The agent causal libertarianism, in spite of advocating a version of dualism, which might not be a desired consequence of solving the free will and determinism problem, has the potential to explain one of the bigger issues that emerges in relation to any event-causal account. One of the important features of free will seems to be the control aspect, where we are free as long as we can choose and control our actions. If we consider events as causes of our actions, it is hard to pinpoint how the control aspect could be realized there. For the proponents of agent causation, the problem of control is easily solved, since the control- unit is the agent himself.

Libertarians and the problem of luck.

Another major issue that libertarians have to provide solutions for is luck, which can influence the control one can have over an action. Mele (2006) provides a clear illustration of the problem through an example of Diana, a goddess with a plan to create a world with a libertarian free will. Diana believes that in the world in which the agent's act freely, they are able to do perform a different action than they actually did, but is her world living up to her expectations?

Her worry, more specifically, is that if the difference between the actual world, in which one of her agents judges it best to A straightaway and then, at t, decides accordingly, and any possible world with the same past up to t and the same laws of nature in which he makes an alternative decision while the judgment persists is just a matter of luck, then he does not freely make that decision in that possible world, W. (Mele, 2006, p.8)

Why does Diana think this way? If the conditions did not change, and the individual's judgement about the best course of action remained the same, it seems that the different action performed by the individual is just a matter of bad (or good) luck. In the end the two possible worlds: one actual and the other one where the individual performed a different action, are exactly the same up to the moment of the action, thus there is no reason for them do lead to two different actions. Moreover, it seems that in the world where the alternative action was performed, if it was in fact a result of bad luck, the individual should not be held responsible for it, and therefore we can question if he in fact executed the action freely (if responsibility is interchangeable with free action). If we then assume that the world is indeterministic, we might encounter a situation where every decision we make is in fact a result of some randomness. The choices that we make are in such a universe a matter of luck, and, as such, most likely not free, or at least once that we are not responsible for.

Hard determinism.

Compatibilism, sometimes also called soft determinism (e.g. Pereboom, 1995), assumes that even in case determinism is the case, or even though it is true, depending on the interpretation, free will and moral responsibility are still possible. Hard determinism, in its hardest form, on the other hand states that:

[…] since determinism is true, we lack the freedom required for moral responsibility, and hence, not only do we never deserve blame, but moreover, no moral principles or values apply to us. (Pereboom, 1995, p.21-22)

Hard determinists are then also, in some sense, incompatibilists, but as libertarians accept the truth of indeterminism, hard determinists claim that determinism is true, and therefore we cannot have the type of free will that we usually consider necessary for moral responsibility. Some of them (e.g. Strawson, 1986, Pereboom, 2001) argue that the discussions on the compatibility of free will with determinism, and the failure to provide a one, consistent account of how free will could work, proves the inconsistency of the concept itself.

Pereboom's own position is a somewhat smoother version of hard determinism, where the fact we do not possess the type of free will that's usually seen as necessary for moral responsibility, does not immediately entail that the moral principles and values are to be rejected or changed. Pereboom does not reject the possibility of indeterminism completely, leaving the possible interpretation of the state of the world as in quantum mechanics open, but he claims that the type of indeterminism is incompatible with free will, at least understood as moral responsibility. Pereboom himself calls his position hard incompatibilism, to distinguish from the usual hard determinism.

In Pereboom's (2001) view the agent can be morally responsible for an action, if he can be blamed or praised for it. The sole fact of the action being performed and authored by a specific agent is enough for this agent to be seen as morally responsible for it. Pereboom's evaluation of agent's action as the basis for moral responsibility seems to be mainly concerned with the facts of a matter rather than any attitudes exhibited by the agent. In his earlier works, he defines being free in the right sense as having

[…] desires that genuinely belong to the agent make up the immediate causal history of the action. (Pereboom, 1995, p.22)

The basis of our responsibility for action lies in our decisions, where the notion of control is central. Action is unfree when:

[…] performed as a result of brainwashing of some types of mental illness. In such cases, desires that genuinely belong to the agent do not play the causal role necessary for the action to be genuinely free. (Pereboom, 1995, p.22)

Pereboom's version of hard incompatibilism can be then characterized by its approach to action where all actions can be divided into two categories: (1) the actions whose causal history is beyond our control, and (2) random events. The latter category can be further divided into: (2.1.) truly random events, where they are not produced by anything, and (2.2.) partially random, where the factors agents are not in control of contribute to them, but not determine them. Since all the events lie on the continuum between the ones having causal factors beyond our control, to the truly random ones, and since free will requires control, Pereboom concludes we do not have free will, since no action is, in fact, in our control.

Folk approaches to the compatibility of free will with determinism.

Since free will is a notion so deeply embedded in our daily practices, with the rise of experimental philosophy, the interest in the folk intuitions regarding some old problems also spiked. As in the debates discussed above the main focus of experimental philosophers was the compatibility of free will with determinism. Two strong possibilities were taken into account: people are either natural compatibilists or to the contrary think that free will is only possible if the world is indeterministic.

The view that people are natural compatibilists was supported by conclusions of the experiments conducted by E. Nahmias and colleagues (2007). It should be noted that in their views free will is bound with moral responsibility, something that can affect the construction of the experimental questions, and through the potential responses of the participants. However, it should also be stated that such move makes sense from methodological point of view, where asking if the person was free might be misunderstood by the participants, while responsibility seems to be a less ambiguous notion, at least in my opinion. Therefore, as I have done throughout, I will use the two terms interchangeably, although I favor separation of the two issues.

The potential ambiguity of the notion of free will can be linked to a more general issue of connecting the terms used in different areas that discuss the same notion and the day to day linguistic practices. The latter connects free will with the “mental” vocabulary, where the explanation is given in terms of desires, decisions and plans. On the other hand, neuroscience has the tendency to provide explanations in terms of physical states and mechanisms, something that is seen as incompatible with the abovementioned mentalistic explanation. The two can be interpreted as incompatible, and the mental is the reduced to the physical or rejected completely. This, as Nahmias and his colleagues recognize, makes investigation of folk intuitions challenging, as for some participants the scientific, mechanistic explanation is simply incompatible with the way we discuss free will in the context of day to day conversations. This in turn might influence the answers given during the experiments, where each scenario presenting the results of neuroscientific investigations will be automatically deemed as unfree.

The folk intuitions can be then seen as incompatible with reductive mechanisms rather than with the deterministic universe, which makes it that much more challenging to obtain a set of precise conditions necessary for free will to be possible.

Based on these issues, Nahmias distinguishes two, potentially independent, issues: (1) the compatibility of free will with mechanistic reduction, or as he calls it mechanistic incompatibilism, it seems that through this Nahmias understands simply providing explanations usually given in mental terms, in mechanistic terms; and (2) the compatibility of free will with determinism, or pure incompatibilism, here determinism is understood again as causal determinism, in the most basic sense (e.g. definition of Roskies discussed in the beginning of this chapter). The reductive mechanisms can be still considered problematic, even if determinism is either not the case, or is not a problem for free will.

The confusion between the two separate problems is what leads other philosophers, as Nahmias argues, to consider people naturally incompatibilistic. When the problems are clarified and distinguishes, it is shown that the folk understanding sees the mechanistic approach, and not determinism per se, problematic. The hypothesis of natural incompatibilistic tendencies can be found, for instance, in papers by Nichols and Knobe (2007), and Nahmias paper (2007) aims to revisit those conclusions.

Based on a series of experiments (e.g. Nahmias, et al. 2005, 2007 Nahmias, 2006), mainly surveys, Nahmias concluded that individuals have compatibilist rather than incompatibilists intuitions. He found that the majority of subjects described individuals taking decisions in deterministic scenarios, as acting out of their own free will and morally responsible for their actions. Besides identifying those intuitions, the experiments helped to established what is usually understood as determinism by non-philosophers. He found that the experimental subjects usually equated determinism with lack of free will, meaning that the usual descriptions of the world explicitly as deterministic would not work, and other ways of describing universe's nature had to be used. That apparent synonymous understanding of the two notions, can be seen as the reason why in other experiments the intuitions are describes as those of incompatibilists.

Different scenarios that were used (Nahmias et al., 2005) included:

(1) A world with the Laplacean notion of determinism, where a supercomputer was able to predict everything that will happen in a future with 100% accuracy. In this world a person was predicted to rob a bank on a specific day which in fact happened. When asked if he was acting out of his free will, the majority of the participants of the experiment said yes. On order to avoid bias, caused by blameworthiness of chosen action, also scenarios where the actions were neural or praiseworthy were chosen. The trials yield similar results in all cases. Similarly, if the question was bout moral responsibility, in place of free action, the results pointed to the individual being morally responsible.

(2) In the second world, the beliefs and values of each individual were presented as caused by the combination of one's genes and environment. In this world two twins were born, and adopted by two different families, one by a selfish, money-loving family, and the other one by kind and honest. Both individuals supposedly differ in upbringing, but are similar in their physical traits, as they share a gene pool. Both of them, at some point in their life find a wallet with money, the selfish raised decides to keep the money while the kindly raised, gives the money back to the owner. Again,

the participants of the experiment were asked if the actions of the twins were out of their own will, and significant majority judged both actions as free.

Again, in the result of both studies suggested that people are natural compatibilists, and their apparent belief in incompatibilism stems from the way determinism is usually understood.

Aside from testing the content of folk intuitions of free will (Nahmias et al., 2005), Nahmias also investigated the features of cognitive processes driving those judgements (Nahmias et al, 2007). The authors hypothesize that the manner of presentation of determinism might influence not only the agent's perception of themselves, but also his assessment of the actions of other individuals, since our attitude toward others hinges on our belief in their free will. In other words, if we can relate to people, and recognize their psychological reasons for making certain decisions, the more likely we are to ascribe free will to them.

In order to investigate this, a number of scenarios were developed, which included (1) neuro scenarios, where actions were described in mechanistic terms, and (2) psychological scenarios, where the descriptions were in psychological terms. In the deterministic possible worlds, agent's behavior was described in terms of sufficient antecedent conditions, which necessarily caused that specific event. The scenarios included and alternative universe that was very similar to Earth (two scenarios), and real world under the assumption of determinism. Again the results of this series of experiments show with statistical significance, that people are not natural incompatibilists, and they do not consider determinism as being a threat to either free will or moral responsibility. Most of the participants found free will less plausible in the scenarios with mechanistic description, and if the descriptions of the agent are abstract, rather than specific (Nahmias, 2007, p.232). The results provide then further reasons to believe that it is the mechanistic stance, rather than determinism, that's the real obstacle for free will, and influences people belief in it.

Moreover, the results strongly suggest that the language used by the researcher has an impact on the way we perceive an agent. A mechanistic terminology, not including concepts like causality of consciousness, or rationality in making choices, seems to lead individuals to rejection of the existence of free will. On the other hand, using the “mentalistic” language is more likely to bring about results suggesting the existence and causality of free will. This dualism of descriptions, influencing the way in which we perceive scenarios is interesting, and might suggest a need for a shift of focus in the research on free will, from the problem of determinism, to the issue of mechanisms.

Empirical testing also led to opposite conclusions, where folk intuitions are shown to be incompatibilists or libertarian (e.g. Nichols, 2006). In a series of experiments, he has presented the participants with variation of scenarios, using descriptions of different possible worlds rather than referring to “our” world.

In the first experiment participants were shown a scenario that included a vast number of possible worlds, each of them inhabited by the same individual, with identical physical make up and upbringing. In one of the worlds the individual decides to learn how to walk the tightrope. The participants are then asked to determine in how many of the other worlds, the same individual would do the same thing. If the subjects replied that it would be all of them, the answer was coded as deterministic. A significant number of participants chose the “all of them” option, leading Nichols to conclude that their intuitions, about the nature of the world, were in fact deterministic. Similar results were obtained in other scenarios used, however the participants varied in their responses depending on the broader context presented in the scenarios. In some of them, individuals will be more prone to deem responsibility as incompatible with determinism, while in others, the nature of the universe does not seem to be an issue.

In other trials Nichols presented the participants with an array of different events, some of them being physical, like a pot of water boiling, while others being choice-events, like choosing a specific type of ice cream rather than another. The subjects were then asked if, assuming that everything leading up to the point of choice or the physical event had been the same, the event had to occur. Through the structure of the experiment Nichols aimed to further investigate the compatibility of determinism with free will, while avoiding openly defining the situation as deterministic. The results of the experiment showed an increased inclination to reject determinism in choice events, than in the physical ones (Nichols, 2006). When presented with two possible worlds, one describes as strictly deterministic, while the other one indeterministic, the participants were more likely to agree that the universe we live in is more like the latter rather than the former. Since the only indeterministic feature of the universe was the possibility of choice, Nichols interpreted it to mean that participants deemed that for a choice to be possible indeterminism had to be the case.

The difference between the two sets of results could be explained by the confusion between the mechanistic description and determinism, as suggested by Nahmias. Another possible explanation involves the role of emotions in our assessment of situations. All experiments discussed here pointed to a role played by the description of both the events (mentalistic vs. mechanistic descriptions), as well as the broader context (our planet vs. imaginary possible planet). Since the participants of the experiments seem to be influenced by both the event description and the context, it leads Nichols to hypothesize that what matters are emotions. In cases when emotional triggers are present in the scenarios, the intuitions of the participants seem to lean towards incompatibilism. In emotionally salient scenarios, people tend to be compatibilists. Emotionally loaded events, which we can relate to, make us believe in free will, while purely physical scenarios push us towards belief in determinism and the falsity of free will.

Alternate Possibilities, Moral Responsibility and Free Will.

The principle of alternate possibilities, on which van Inwangen's consequence argument was based, puts strict conditions on free will as it assumes that, in order for an agent to be considered responsible for his actions, and therefore free, he needs to be able to have done otherwise than what he had in fact do. Combining this premise with the deterministic nature of the world, we are bound to ask ourselves: In the world where everything is determined by the totality of the previous states combined with the laws of nature, could anything else than what in fact happened happen? But is it necessary to be able to do otherwise to have free will, or to be morally responsible? Since most of the discussions on the principle of alternate possibilities refer to moral responsibility, I will once again treat it as synonymous with free will.

Interestingly, while the compatibility of free will with determinism is, as we have seen in the previous parts of this chapter, often questioned, very few seem to be as inclined to question and/or reject the alternative possibilities and their role in assessing agent's responsibility. Until Frankfurt, it was usually assumed that the existence of actual possibilities is necessary, thus the problem of determinism, he however questioned if those possibilities have to truly exist 'out there', or if it is sufficient that we have an illusion of choice.

The compatibility of free will with determinism is often questioned and broadly discussed, yet very few seem to be strongly inclined to reject the alternative possibilities principle and its role in assessing agent's responsibility. Frankfurt argues that the principle

[…] has generally seemed so overwhelmingly plausible that some philosophers have even characterized it as an a priori truth. People whose accounts of free will or of moral responsibility are radically at odds evidently find it a firm and convenient common ground upon which they can profitable take their opposing stands. (Frankfurt, 2013, p.139)

Despite the broad approval of the principle, in spite of the issues that it entails, Frankfurt argues that the person could still be considered morally responsible, even if the principle itself was false. In other words, we could still be deemed morally responsible for an action, even if it was not, at any point, possible for us to do anything else.

We generally tend to believe that all human action falls under the principle's range, but in fact not all action meets said criterion. In cases where outside circumstances play a significant role in agent's choices, such as manipulation of that agene or hypnotic suggestion, the principle does not apply. In such cases the principle is not falsified (as the agent could have potentially done otherwise) but, at the same time, we would not see the agent as responsible for his action. Being under influence, manipulation or coercion, relieves us from being responsible, but, as Frankfurt argues, it is not because of the principle not being effective, in the end it is also not falsified, but because of other conditions. To support his thesis Frankfurt offers several examples in which the individual, although coerced and manipulated, also decided to execute the same action prior to the manipulation and coercion. To illustrate this, we can think about an evil scientist that implemented a chip in our mind, which, when triggered will make us kill the neighbor's cat. However, without his knowledge, we have previously already decided to kill our neighbor's cat, and we would have done so regardless of the mad scientist's manipulation. It seems that in those cases, the agent could still be considered morally responsible, which would entail that it is the agent's reasons, and not his ability to do otherwise, play a role in assessing his moral responsibility for an action.

Based on those conclusions, Frankfurt proposed a reformulated version of the principle where: *[…] a person is not morally responsible for what he has done, if he did it only because he could not have done otherwise"* (Frankfurt, 2013, p.147).

The second principle emphasizes the agent's guidance control (as opposite to the regulative control found in the first version of the principle) over his action. In the cases where an agent had reasons to execute an action, based on the revised principle, we have good reasons to think that he would still be morally responsible for an action, even if he could not have done otherwise. This brings into question the necessity of actually having different action-alternatives in deeming and action free. Frankfurt additionally argues that this new version of the principle does not stand in opposition to the deterministic conditions.

Is the principle of alternate possibilities a true threat to free will, given the deterministic world? Some (e.g. Dennett, 1984) argue that the principle is in fact required for the agent to be considered responsible, since we do not necessarily exempt people from responsibility based on their inability to do something else. When discussing responsibility we, more frequently, cite reasons for action, rather than other alternatives that might have been exercised. We will hold people responsible when they stand by their actions, regardless if they could have done something else or not, therefore the central premise of the principle, that of alternative possibilities being necessary for moral responsibility seems false. If then we consider free will as synonymous with, or at least having the same basic necessary conditions as, moral responsibility, Frankfurt's argument seems to suggest that what is necessary for free will is not that we can do otherwise, but rather what is the causal origin of our actions. Then actions are free when they are caused by, somehow, our reasons.

Is it possible to change the laws and the past?

A different way to deal with the issues stemming from the combination of the principle of alternate possibilities, and determinism, which led to the consequence argument, was proposed by David Lewis (1981), who revisits the question of the unbreakable nature of the laws of nature. Although he sustains that the laws of nature in the strong sense remain unbreakable, stating that:

[…] a genuine law is at least an absolutely unbroken regularity (Lewis, 1981, p.292),

this however does not imply that all the instances are governed by laws, which could possibly only supervene on facts.

Lewis thinks that in order to be able to do something else than what we did, we need to be able to break laws of nature, since any action if determinism is the case is a combination of the previous state of the universe and the laws of nature. Since we cannot change the previous state of the universe, something about the laws of nature would have to be changed to allow the other possible action to take place. He argues that, although breaking laws of nature in a strong sense is not possible, it is doable in a weak sense, which would require that some law, would have to have been different. The change in the law is not up to the agent, therefore we are not the ones that "break" the laws, but rather the law was already different when we started acting, and that led to a different action than the one we are considering. This solution of course heavily relies on specific interpretations of being able to do otherwise, as well as the notion of breaking the law of nature. The latter is seen as ambiguous in the Consequence Argument by e.g. Helen Beebee (2013).

Helen Beebee (2013) clarifies the problem here, by offering an example of Sue who is making a cup of tea.

So, if Sue's making a cup of tea, say, is entailed by the laws plus the past, it doesn't follow that Sue could not do otherwise than make the tea, in a perfectly good sense of 'could have done otherwise. (Beebee, 2013, p.58)

In such cases breaking the laws of nature, as required by the Consequence Argument, can mean either of two things:

(1) Not being able to do or cause anything that could break the laws of nature (in this case the agent would have to break the laws in the strong sense)
(2) Not being able to do anything such that, if something else was in fact done, some law of nature would have to be different (breaking the law in the weak sense, where the condition of being able to do otherwise relies on the existence of different laws of nature than the ones that currently exits).

Nobody has the power to break the laws in the first sense, since understood as such they are a priori truth and simply unbreakable. However, it seems that the second one is perfectly plausible, and compatible with determinism. Moreover, we seem to often use this understanding in our daily life, where we make counterfactual claims of the sort: If the bus would not have been late, I would have arrived on time.

Lewis' approach, although having the potential to save the principle of alternate possibilities, seems unconvincing to incompatibilists, who, as Kane discusses, could simply refute the distinction between the weak and strong sense of breaking the laws of nature. Kane argues that, although the division onto the two senses is valid, the Consequence Argument clearly commits us to the stronger reading, and therefore Lewis' solution does not apply, when he states:

We incompatibilists do understand Lewis's distinction between weak and strong senses, but having taken note of this distinction, it is clear to us that the proper reading of the premises of the Consequence Argument is the strong sense of 'can render false'. The only think that could recommend the weak sense to anyone, as we see it, is a prior commitment to compatibilism. (Kane, 1996,p.50)

Although not a conclusive solution and one might add such probably requires a conclusive answer as to what the nature of the universe actually is, Lewisian approach is promising. It shows that, depending on the understanding of the "can" notion, compatibilists could have reasons to show that alternate possibilities are in fact possible, and therefore render the Consequence Argument false by rejecting its main principle.

Conclusions

This chapter was focused on two interconnected problems, usually at the core of the philosophical debate on free will, the problem of determinism and, linked with it, the problem of alternate possibilities.

The possible arguments, in connection to determinism problem, can be divided into three main categories: (1) compatibilism- there is no tension between free will and determinism; (2) incompatibilism- for free will to exist the universe should be indeterministic, with libertarians arguing this is in fact the case, and (3) hard determinism- the universe is deterministic, we have no choice because nothing is up to us, free will does not exist. As we have seen all of the arguments have their shortcomings and the answer to the question of compatibility lies in the proof of the actual nature of the universe. The positions are very clearly illustrated by Roskies:

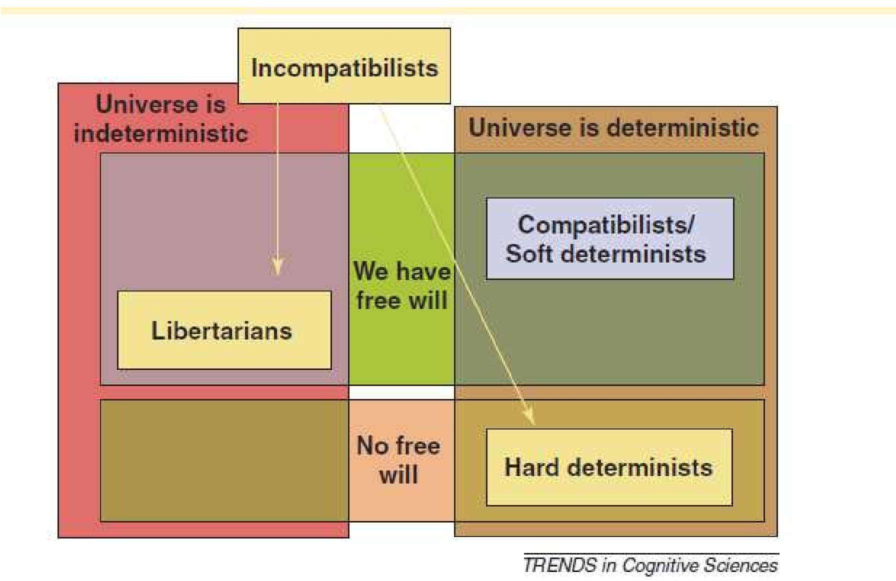


Figure 1. Compatibility between free will and the state of universe (Source: Roskies, A., 2006, p.420)

Additionally, the folk intuitions regarding the compatibility were discussed. Also here no determinate answers were found, with the results obtained supporting both incompatibilist as well as compatibilist intuitions. The contradictory results can be due to (1) misunderstandings in defining what determinism is, and possibly confusing it with the mechanistic explanation, the latter usually seen as incompatible with free will, or (2) the emotional charge of the scenario presented, where in emotionally neutral scenarios the participants tend to have compatibilist and in emotional scenario incompatibilist intuitions.

In the second part of the chapter I have shown Frankfurt's approach in denying that the principle of alternate possibilities is necessary for an account of moral responsibility, and from that of free will. I have also discussed an alternative route taken by Lewis is analyzing what kind of conditions would have to be fulfilled for alternative action to be possible at al.

I have shown that the internal conditions for the possibility of free will rely heavily on the, up to date, unanswered problem of the nature of the universe, which some take as a fact, while others as two equally valid metaphysical problems. However it seems that in all discussions, especially visible in Frankfurt's refutation or reformulation of the problem of alternate possibilities, what is often overlooked is the apparent importance played by the way an action is caused. Frankfurt's argument hinges on the existence of reasons that play a causal in the action generation, while other discussions often point to the existence of a decision, or intention that has an impact on the action generation. This issue, what kind of causes should an action have to be considered free, are the central issue for another type of discussions, based on the experiments conducted in neuroscience.

Most important points of this chapter, including my position:

1. As I have stated before I remain agnostic on the compatibility between free will and deterministic nature of the world. I however think that the problem cannot be understood fully without the arguments discussed from the two, most popular angles, presented here.
2. The convergence points between the three different axis of research in philosophy are: causality and the types of events that cause actions that can be defined as free. Therefore, to provide an account on which free will is possible, it is necessary to decide both what we consider causality, and what are the various possible causes, that are the source of free action. This will be discussed in Chapter Six.
3. Folk discussions on free will point out another important issue: the incompatibility of free will with mechanistic explanation. This combines two problems: the relationship between causality and explanation, and the approach to causality that can work in

the discussions on free will. The problem will be further elaborated in the next chapters.

Chapter 3. Neuroscience and the free will problem.

With the recent advancements in brain imaging techniques, the interest in the questions such as free will and consciousness increased also in the scientific domains. First experiments on the causal origin of our actions were conducted in the sixties, and since then the interest has grown significantly. In this chapter I will analyze the neuroscientific experiments attempting to find empirical proof on the existence of free will. I will begin by discussing what are the concepts neuroscientists operate with, starting from the voluntary action itself, which seems to be synonymous with free action. I will then proceed to discuss the most influential research, done by the pioneers of the field: Libet and Wegner, focusing mainly on the former. I will present some of the issues that have been identified in connection to both, (1) the assumptions of the experiments as well as (2) its conclusions, both by scientists and philosophers. The main goal of this chapter is to identify the, often implicit, assumptions of the experiments, both regarding the nature of causality, as well as the events or properties that take part in such relationships

What is free will for neuroscientists? The concept of voluntary action.

As we have seen in previous chapters the notion of free will can have multiple interpretations, depending on the setting of the problem against facts about the world and the individual. In this chapter I will turn my attention to the problem of free will, as it is discussed in neuroscience, and in neuroscientific experiments.

Since the experiments are focused on free will "in action" the latter is usually defined through voluntary action. In this subsection, I will provide examples of both explicit definitions of voluntary actions, as well as what is implicitly assumed about them, based on the construction of the experiments. I will focus mainly on Libet-style experiments, and, after describing the initial assumptions and the methods used, I will present the most common points of criticism of those experiments. I will argue that based on the many results, we have good reasons to believe that Libet's do not in fact threaten free will. I will also, although more briefly, discuss the results of Daniel Wegner, who, arrived at similar conclusions as Libet, but based them not on empirical experiments on subjects, but rather on the study of automatons. His position on free will is also stricter than that of Libet, since he rejects the existence of free will in any capacity, while Libet tries to preserve it as a veto power.

Because of their approach to the topic, the neuroscientific experiments have more in common with the problems discussed in the philosophy of action, and the discussion on the potential causes of human action, rather than with the classical debate on free will. The issue of free will is here, almost exclusively, discussed in connection to action initiation with some (e.g Haggard) explicitly expressing their belief that the problem of free will is in fact the problem of the causal origin of our motor actions:

[…] I will consider free will from the perspective of the human motor system. This view is justified because the problem of free will is a problem about the causes of our motor action. […]. The approach from motor neuroscience begins with the fact of the motor action, and then traced back to identify the particular stages of information processing in the brain that eventually produce it. (Haggard, 2011, p.8)

Haggard's approach emphasizes an important methodological feature that accompanied the research on free will, as conducted in neuroscience, namely the necessity of the overt motor action to occur. The specificity of research in this domain is marked by its need for something measurable, like an action or behavior, which is then used as a starting point for the analysis of the brain process that occurred right before it. The measurements are done by backtracking from the action to its potential causes, and in free will those include the ephemeral intentions and the deeply physical brain activation. This type of approach is also used by Libet, whose experiments will be discussed in detail later on, where the overt action serves as a starting point for the researcher to check what happened before the movement in the participant's brain. As we will see, there were some attempts to investigate decisions that did not result in movement (like vetoing the action already underway), but usually their results are mapped onto the results of tests where the overt movement did occur, to have a better idea of the timeline.

The issues examined by neuroscientists are what I call the internal conditions for free will, examining what could cause agent's behavior, not from the perspective of the broader context of the action, but rather what it is inside the agent that is the causal origin of the action. The connection between the experimental approach and the classical philosophical debates on the topic is less obvious. It seems that beyond the main focus, the free will, the question of determinism is often dismissed, e.g. Haggard (2011) argues that determinism is a fact, and any account of free will to be plausible, has to be compatible with it. Others, like Libet (2004), admit that both determinism, as well as indeterminism are only theories, and both are still possible.

Focusing on the causes of our actions brings a new set of problems. One potential complexity has to do with the complexity of the causes that are involved, as well as the flexibility of the individual's decision (e.g. Haggard, 2014). It is not necessarily clear what is being discussed under the "causal chain" umbrella, the various possible events that result in action. Said chain might be seen as short, where only the most proximal events are taken to be causes, and the causes of those events are not included in our investigation. It can also be assumed that the chain includes also the causes of those events immediately preceding the action, possibly both conscious and non-conscious, that all have some sort of influence on the final outcome. Pinpointing to a specific one cause might be difficult, when we assume that a specific movement is a result of a long process involving both mental and physical.

For the experiments, because of the technical challenges, the only processes that, at least initially, were testable, were the ones that result in an overt action. There are good reasons to think that, at least some, deliberations and decisions involve not acting, or changing our minds midway through the action execution, but linking voluntary action with an overt action has some additional advantages here. Libet defines the process taking part in the shaping of the final act as including:

[…] consciously deliberating alternative choices as to what to do and when, whether or not to act, whether or not to comply with external orders or instructions to act and so on. If any of these processes are to result in the motor performance of a voluntary act, they must somehow work their way into a 'final common motor action pathway' in the brain. Without an overt motor performance any volitional deliberation, choosing or planning may be

interesting for its mental or psychological content but it does not constitute voluntary action. (Libet, 1985, p. 530)

The voluntary action is therefore necessarily an overt action, and moreover, the various steps of deliberative action have to somehow connect and together, through the *final common motor action pathway* cause the movement (or muscle contractions).

Since free will is exercised in the voluntary action, the definition of the latter marks the characteristics of a free action. Again, the definition of either of the concepts differs between different authors, with Libet defining voluntary action as combination of three conditions, namely:

(a) it arises endogenously, not in direct response to an external stimulus or cue; (b) there are no externally imposed restrictions or compulsions that directly or immediately control the subject's initiation and performance of the act; and (c) most important, subjects feel introspectively that they are performing the act on their own initiative and that they are free to start or not to start the act as they wish. (Libet, 1985, p.529-530)

I understand the first condition as excluding reflexes from the pool of the potentially voluntary actions, they arise in a direct response to an external stimulus, e.g. you will move your leg when somebody hits the right spot in your knee. If we think of this condition in terms of dual process theory, dividing human behavior into two types: (1) stimulus-driven behavior, a direct result of an external stimulus, often equated with an automatic one, and (2) goal-directed behavior, which is often seen as conscious, slow, and non-automatic, Using this framework, the actions that Libet would be interested in would be the second type only. This condition alone, without takin other two into account, could however be seen as including some automatic action, if we limit the understanding of the sense of the notion response to the external cue (here I will however assume that the only interesting, for Libet, types of behavior would be the non-automatic one).

The second condition excludes compulsions and things the agent might be, in one way or another, forced to do as any of those would limit agent's ability to control the initiation and the performance of the act. Again, it should be noted, that the conditions Libet is describing here are not aimed at explaining what free will is, but rather describe the conditions which, when fulfilled, allow us to recognize an action that is a result of the exercise of free will.

Third condition is that an agent recognizes himself as the causal source of their actions, and they have the ability to start and veto the action at the moment of their choosing. This condition presumes some form of a sense of agency that the individual has, while initiating the act. Libet (1985) emphasizes the importance of the third condition using an example of a patient who had their motor cortex stimulated, which led to the individual performing a movement. In all instances, the patients felt that the movement was something done to them rather than by them, the sense of agency was missing and the movement could not be considered voluntary (although here it seems that the first and the second conditions are also violated).

Those three conditions are then used as a blueprint to identify voluntary actions, and mark the explicit assumptions Libet makes in connection to free action. Another set, the implicit assumptions, are probably what is more important here, and are suggested by the construction of the experiments. The way the experiments are conducted entails the existence, and possible competition, of two, seemingly independent, causes of the action: (1) the physical brain activity of some sort, and (2) the, presumably mental, conscious decision. In order for an action to be voluntary it has to be initiated by the latter rather than the former. This condition, implicit in the papers detailing his experiments (Libet, 1982, 1985) was made explicit later on (Libet, 2004) and it stems from his experimental assumptions: the conscious will has to precede the brain processes that then lead to the act, or it has to appear simultaneously with them. The connection between the two is not further explored, which could suggest some form of dualism, since, significantly, the conscious decision seems to appear out of nowhere, not connected to the previous brain activation.

A somewhat different definition of voluntary action is given by Haggard (2011), who conducted a series of modified versions of Libet's original experiment. Unlike for Libet, for Haggard the action to be considered free has to have two characteristic features.

The first is the well-known 'Could have done otherwise' feature. In philosophy, this is often used as a test of compatibility with determinism: given the preexisting state of the universe, could the action have occurred differently or not occurred at all? Movement neuroscience *understands 'could have done otherwise' in terms of a process of action selection, that arbitrates between competing action alternatives, and selects the one that eventually occurs. […] if we have evidence from brain measurements that two different action alternatives were developed in parallel, up to a late stage just before action execution, and if we can identify the particular brain process that led to one action alternative being selected and the other deselected, we may say that the person, or animal, could have done otherwise.* (Haggard, 2011, p.9).

Here, unlike in Libet's definition, some of the problems so broadly covered in the philosophical debates on the free will problem, can be found. Haggard refers specifically to the principle of alternate possibilities, which challenges if, under deterministic scenario, free will can be possible, as we do not seem to have a choice. However, Haggard's understanding of the alternate possibilities differs from the one usually found in philosophy. Rather than actually different possible actions, I assume, existing actually in the world, Haggard refers to different action-alternatives being processed by an individual. As such his approach seems closer to the one by Frankfurt, where the different options do not have to actually exist and the deliberation and reasons for action are what is important for the assessment of the action. This approach seems to, moreover, be compatible with both indeterminism and determinism, as it seems plausible that the state of the universe was such that it deterministically supported the development of two action alternatives up to the moment of the, deterministic, choice between them. However, the moment of choice remains problematic, as it is unclear what, or which brain process would have this sort of control and the ability to choose between them? It is not clear. It seems that Haggard attributes the properties we usually attribute to the agent, like intention or decision, to a neuron or a cluster of neurons, or some physical process. Later in the same publication Haggard describes the selection of the action as being done by a mechanism in charge of eliminating the additional possibilities, but again this type of control is usually attributed to a conscious agent.

The second requirement is that:

[…] there must be a 'I' who could have done otherwise. Moreover, the agent must be conscious that they are making the action. We will therefore refer to this as the 'subjective

agency'. Therefore, concepts of free will do not apply when this subjectivity is absent, for example to a person that acts in their sleep […]. Neuroscience treats 'I' as being synonymous with an individual's brain. Therefore, it approaches the awareness of action by distinguishing between those action-related processes in the brain that typically generate conscious experience, and those that do not. (Haggard, 2011, p.9)

Similarly to the third condition described by Libet, also Haggard seems to think that the agent has to recognize his own agency if the action is to be defined as voluntary. The agent has to see himself as the causal source of the action.

As for the implicit assumptions, since Haggard performed and array of, often modified, Libet-style experiments, it is plausible to think that he accepts the division onto two potential causes: the physical brain activation on the one hand, and the mental conscious decision on the other. However from the abovementioned quote, it seems that the conscious awareness of the action is generated along other action related physical processes. It is to some extent unclear if he just assumes that conscious awareness of the fact that we are an agent is connected to other processes that end in action generation, or rather that the whole conscious decision is generated in that way. The latter understanding would lead to the contradiction in the assumptions of the experiment, from one we are hypothesizing that it is possible the conscious decision has a causal impact. Based on the other, we would be committed to saying that such conscious decision is a result of action generating processes already underway, before we started the experiments. Here I assume that Haggard just means the awareness of the action is somehow connected to the processes of action generation (the question of what that potential connection would be, will be explored further in the next part of this dissertation) which would not exclude the potential causality of the conscious decision.

Similarly, as in Libet's case, Haggard does not investigate the connections between the mental and the physical, and does not clarify further how the two could be related. However, his second requirement could be red as accepting some form of mind-brain identity theory, which, as we will see further in this dissertation, could be itself problematic.

In the next sections I will provide a detailed description of Libet's experiments, and I will discuss what kind of intuitions about both causality, as well as the properties of events that enter causal relationships, Libet seems to exhibit. I will then discuss a variety of problems with his approach, from various angles: neuroscientific (what the readiness potential in fact is, does it represent an action initiation, other preparatory processes or is just a brain noise), and philosophical (does the results of the experiments really threaten the existence of free will). I will show that both perspectives provide us ample reasons not to think that Libet's experiments truly show something conclusive about free will. In the process I will describe some of the modified versions of the experiment, which, although often problematic for the same reasons as the original experiment, provide us with more insight on how the action arises.

Benjamin Libet and the surge of the experimental interest.

Benjamin Libet's experiments are currently taken to be the classic example and a blueprint for the newer investigation into the free will problem. His research focused on the relation between a specific brain activation, the readiness potential, and the overt action performed by an individual. The first discovery of the readiness potential is due to experiments by Kornhuber and Deecke, and Libet modified their original approach. He focused on the causal source of a voluntary, spontaneous action of individuals, and through EEG studies tried to determine if what starts our action is in fact the physical process or rather the conscious decision. His experiments have become a standard in neuroscience, and many followed, conducting modified versions of Libet's study (e.g. Haggard and Eimer, 1999; Soon at al., 2008; Schlegel et al., 2013, 2015; Alexander et al. 2016;), reaching a variety of conclusions.

The readiness potential.

Readiness potential was first discovered by Kornhuber and Deecke (1965) through recordings in motor cortex and the supplementary motor area (similar approach was afterwards taken by Libet who also recorded in those areas). Their initial goal was to investigate human will and the causal source of voluntary actions. Deecke would later write:

We describer a novel method, reverse averaging, from recording brain electrical activity prior to voluntary movements in humans by noninvasive means and presented the first fundamental results obtained with this method. We found that a negative electrical cortical potential consistently preceded voluntary movement and named it Bereitschaftspotential (BP) or readiness potential. (Deecke, 2014, p.1)

The method used by Kornhuber and Deecke, back-averaging from the movement, provided the necessary muscular activation acting as a reference point, from which the preceding brain processes were back-traced. The same method was later used by Liber, and, in a nutshell, the researcher would record brain activity that took place prior to making a movement, and use the movement itself as a mark from which the preceding activity was analyzed, and the links between consistent brain activations that preceded each movement were discovered. Without the movement, those specific preparatory processes would have been difficult to distinguish. As we will see in the proceeding chapters this approach will be one of the sources of criticism of his experiments.

The main difference between the original experiment by Kornhuber and Deecke, and the one conducted by Libet was in the actions used. Libet chose self-paced actions, while Kornhuber and Deecke: fully spontaneous.

Libet's experiment.

For Libet, the initial inspiration to conduct another version of Kornhuber and Deecke's experiment was his conviction that the previous experiments did not investigate truly voluntary actions. He thought that the way the experiment was designed might have compromised the nature of the acts, and because of that they could no longer be considered voluntary. The necessity to repeat the same action hundreds of times, combined with simple boredom the subjects experienced, might have acted as an external influence on the participant's choices. Additionally, Libet (Libet et al., 1982) argued, the RP recorded by Kornhuber and Deecke contained in fact two components instead just one: *the earlier ramp-like component* (Libet et al. 1982, p.322) and a second, more asymmetrical one. The different timing features (one was observed earlier than the other) along with the possibility of their different neuronal origin raised questions about their involvement in the initiation of the voluntary action.

The most problematic features of the original experiment are described as:

(a) Recording an RP requires averaging many events. When these self-paced acts are repeated in a continuous series, with irregular intervening intervals of 3-6 sec are selected by the subject, they become boring and may come to be performed in a stereotypes and almost automatic way, with no assurance that conscious control could be exercised in each trial,

(b) Since subjects were asked to act within an allotted time interval, they may be under pressure consciously or unconsciously to plan to act within the time limit; that is the subject's voluntary choice of when to act may be compromised by an external requirement,

(c) Subjects are required not to blink until just after each act. The need to blink may impel the subject to act, thus serving as an external controlling factor. (Libet, 1985, p.530)

In order to avoid same issues in his updated version of the experiment, Libet aimed to both: (1) differentiate between fully spontaneous and pre-planned movements, and (2) provide instructions that would assure, to the extent it is possible, or at least support full spontaneity in the choice of the moment of movement.

Each trial (out of 40) was treated as a separate event, and the gaps between the individual trials were determined by the participants and dependent only on his or her own readiness to proceed, which would eliminate the boredom factor. There was no time limit for when the participant was expected to perform the movement, which was supposed to encourage spontaneity of the movements, and individuals were allowed to blink if necessary. The results obtained were divided into two different types with the possibility to further divide them into three.

*Type I had an early onset at about -1050 +/- 175 msec and long ramp- like form, resembling self-paced RPs. In type II the main negative shift began at about -575 +/- 150 msec, and at about -240+/-50 msec in type III. (*Libet et al., 1982, p.322*)*

The results of those test, and especially the “long” (in neuroscientific terms) lag between the cerebral processes and the overt action, prompted Libet to investigate the connection between said processes and conscious intention. The initial goal was to check if the onset of conscious decision also appears before the action with such a time lag. The results of the experiments were staggering, and had impacted the way neuroscience thinks about free will ever since.

In order to have a measurable account of the moment when the conscious decision is made Libet (Libet et al. 1983) asked the participants to execute an overt action (move their fingers or flex their wrist) while looking at a *center of the 5 inch circular screen of a cathode ray oscilloscope […] positioned about 1,95 m away […]* (Libet et al. 1983). The spot of light of the oscilloscope was rotating clockwise, starting at the position corresponding to 12 on a standard clock. The participants were instructed to, at the moment of their choosing, perform an action (move their fingers or flex their wrist).

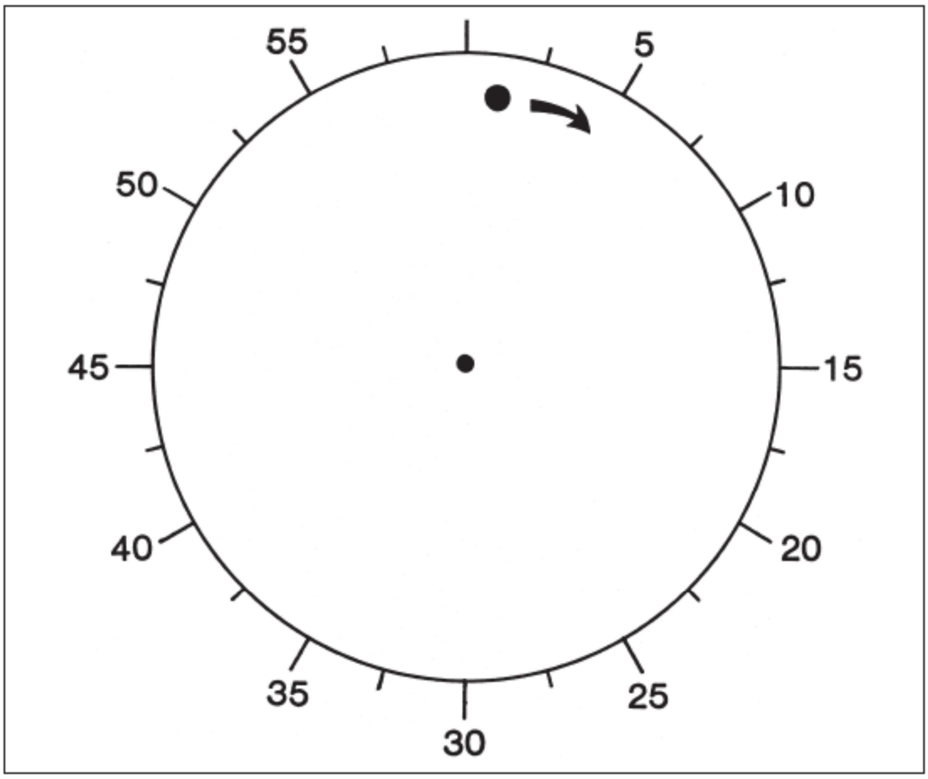


Figure 2. The oscilloscope used by Libet (Libet, 1999, p.50)

On the figure above we can see the design of the original oscilloscope used by Libet. The marks represent ‘seconds’ which translated into 43 sec. of actual time. The spot of light, here marked as a bigger black dot, has been moving around clock-wise. The participant was observing the rotating light, and was asked to remember its position when he made the decision to move.

A lot of freedom was given to the participants as to when the movement had to be executed (as described above, they could take as much time as they needed both during one trial as well as between the trials, they were also allowed to blink). The importance of the free choice in terms of the moment of the action rather than the type of the action was further confirmed by Libet in his later articles (Libet, 1985), where he stated it was the crucial element of his study as it identified the acts as voluntary.

The experiment run in two series, in the first one *self-initiated voluntary acts* (Libet, et al. 1982, p.625) were studied, and the subjects were asked to wait one full rotation of the clock before moving. After that initial rotation they could move at the moment of their choosing, with no time- restrictions. For this specific trial, the subjects were asked not to plan any action in advance, to allow full spontaneity of the decision and the action.

The second series *skin-stimuli at unknown times* (Libet et al. 1982, p.625) the subjects were expected to receive a pulse on the back of their right hand. The pulse was deliver at random, unknown times after the first full rotation of the clock, and their time of delivery was in fact overlapped with the results from the first series and agreed with the perceived moments of conscious decision from the precious series. The skin stimulus series was used as a control group for the measurement of the subjective timing of the intention to act, since the participants were asked to simultaneously observe both the said intention as well as the clock position at the time, which could potentially lead to errors.

The subjects were asked to remember the exact position of the spot of light on the rotating clock, like on a standard watch, when the decision or conscious intention to move first occurred. Then, only after the experiment, they were asked to report that moment, otherwise known as the W judgement.

There were two modes, designed to help the participants remember the exact time of the intention to move: mode (A) or the absolute mode, and mode (O), or the order mode. In the absolute mode the participant was asked to look back on the time scale, on which the rotating light was projected, during the experiment, and report the position of the clock using seconds (the disc was divided into scale that resembled 5 seconds intervals, which corresponded to about 42,7 ms). In the order mode the participants were asked to report the order of the final position of the clock, relative to the recalled position, when their conscious awareness was first spotted.

The notions such as conscious decision, conscious awareness and the onset of intention are often used by Libet equivalently, which can contribute to the confusions in interpreting his conclusions. For this specific section I have adopted the same tactics, however further on we will see that using notions, often not synonymous, as such, is one of the major issues with Libet's approach, as identified by the critics.

The recording and averaging methods were similar for experiments described in, among other articles (Libet et al. 1982), all done using EEG machine. The signal was recorded while subjects were participating in the experiments, and then the results were analyzed using the recorded movement as a reference point. The reported moment, in which the conscious intention to move appeared (otherwise called W judgement, the moment in which the participant first felt the urge or decision to act), was then compared with the EEG readings to establish if there was any brain activity preceding the reported urge.

What in fact was being recorded? Libet asked the participants to time three different subjective experiences (W, M and S), and each series of 40 trials was concentrated on only one of the three for all 40 trials:

(1) In the W series Libet asked the subjects to monitor what he called the onset of the *conscious awareness of wanting to perform a given self- initiated movement* (Libet ety al. 1982, p.627). This specific activity was described as urge, intention or decision to move, and the terms seemed to be used as synonyms, although, Libet emphasizes, the subjects usually settled to using either the 'wanting' or an 'urge to move'. The participants were also asked if at any point, they moved in spite of not feeling any urge, just to follow instructions, which would mean that there were movements without a conscious intention to act. The number of those instances was small, still it indicated that the actions were a result of endogenous experience, and not induced by instruction.

(2) In the M series the participants were asked to monitor the awareness of the actual movement. This type of experience would be related but not the same as the W experiences.
(3) In the S series the participants were asked to track the awareness of sensation, triggered by the random occurrence of the pulse.

Libet described the results of the trials where:

These self-initiated, endogenous acts were indeed found to be preceded by RP's (Libet et al., 1982). When all 40 self-initiated acts in an averaging series were performed with this spontaneous ade lib timing, with no reports of specific preplanning to act, the recordable averaged RP generally had an onset for its main negative rise at about 550 (+/- 150) ms before the motor act began; these were called 'type II' RP's […] (Libet, 1985, p.532)

The reported time of W judgment, on average, preceded the actual movement by 200 ms, yet in all the subjects a visible activation of the brain (the RP) was observed much earlier. The results were then divided based on different types of acts, and so for the fully spontaneous ones, without any reported pre-planning, the RP (referred to as type II RP) was observed already 500 ms before the movement. The muscular activity, recorded based on the signal of the onset of electromyogram, provided the timing of the onset of the movement, and the zero point from which the brain activity was backtracked.

Despite the heavy encouragement of full spontaneity, some pre-planning was unavoidable. These trials were distinguished, from the ones with no preplanning, during the debriefing, after each series of 40 trials (Libet, 1985). In the trails with some pre-planning reported, the difference between the decision and the RP (called type I RP) was even greater, averaging about 800 ms prior to the movement. The W judgments' timing was roughly the same for all of the trials, both with and without reported pre-planning.

In the trials were the skin stimulus was delivered, some error in reporting was recorded, where:

The skin sensations were commonly reported to occur somewhat in advance of (negative to) the actual delivery time […] the amount of error found in the stimulus series did not *qualitatively alter the difference between the onset of RP and W; in fact it generally enlarged the difference.* (Libet, 1985, p.534)

Those results, taken together, led Libet to one conclusion:

It is clear that neuronal processes that precede a self-initiated voluntary action, as reflected in the readiness-potential, generally begin substantially before the reported appearance of conscious intention to perform that specific act. This temporal difference of several hundreds of milliseconds appeared fairly consistently regardless of which of the available criteria for the onset of RP or for the time of awareness are adopted. (Libet et al., 1982, pp.635-636)

The result can then be represented as:

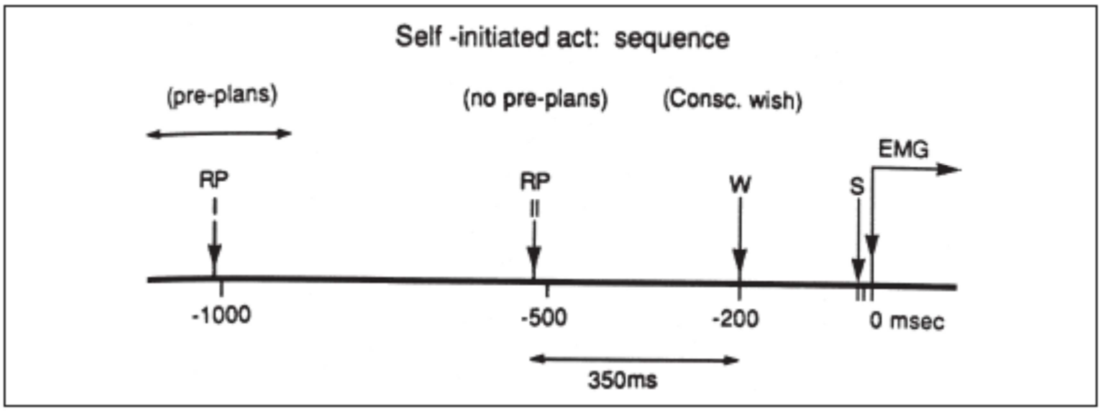


Figure 3. Sequence of events preceding self-initiated voluntary acts (Libet, 1999, p.51)

As we can see on the figure above, there was an observable difference in the appearance of the readiness potential in the instances that included some pre-planning, opposite to action in which the participant of the experiment did not pre-plan. The conscious decision appeared consistently around the same time. The awareness of the randomly delivered skin stimuli (marked S) averaged about 50 ms. earlier than the actual delivery time.

Those results were then concluded using the famous phrase:

[…] the brain evidently 'decides' to initiate or, at least, prepare to initiate the act at a time before there is any reportable subjective awareness that such a decision has taken place. (Libet et al., 1982, p. 640).

The conclusion entailed that, although free will might exist, it is not exercised in the action initiation but rather is a form of control over the action already underway. Libet rejected the hypothesis that RP might not indicate the initiation of voluntary movement but might rather be a preprogramed brain process occurring in the brain at random intervals as not supported by evidence (we will see later on that such hypothesis might have some standing after all). He however did not completely falsify the effectiveness of will, rather argues that instead of being the triggering force, free will can be understood as a form of a veto power. Instead of initiating, our conscious decision can stop the action already under way. This hypothesis was supported by two kinds of evidence from his experiments:

(a) *Subjects in our study of RPs and conscious timings reported that during some of the trials a recallable conscious urge to act appeared but was "aborted" or somehow suppressed before any actual movement occurred […]* (Libet, 1985, p.538)

Here, the participants experienced some form of urge to move, but then aborted the plan, an ultimately did not move, which could serve as an example of a direct control through the veto power. However, it is also possible that not only the movement, but rather both the movement and the conscious decision were somehow aborted. Moreover, given the methodological difficulties, further investigation of this type of experience would be difficult on Libet's approach, as it relies on the existence of some measurable overt behavior.

The second type of evidence was recorded in the M-trials of the experiments where:

(b) *In series of acts to be performed at prearranged times, subjects were instructed in advance to veto the developing intention/preparation to act […] ramp-like event potential was still recorded during >1 sec before the preset time, even though no muscle activation occurred.* (Libet, 1985, p.538)

Here, unlike in the previous example, the participants were clearly asked to veto the action. The type of activation found was similar to the one recorded in the trials with reported pre-planning (hence the long gap between the RP and the moment when the movement would normally occur). This would suggest that some form of veto was exercised over an action. This approach can also be seen as problematic, since, as Mele (2009) argues, it would require that at the initial moment the participant has both the intention to move and the intention not to move. The readiness potential that resembles the one recorded in pre-planning, might be indicative of planning to not move, in accordance with the instructions, and cannot be then considered an example of the discussed veto power.

A number of issues can be immediately recognized both in connection to the actions chosen by Libet, as well as his initial assumptions. Two things draw immediate attention, namely (1) the type of decisions, where they are expected to be fast paced and spontaneous, and (2) the mysterious nature of the conscious decision. The action chosen: fast paced and spontaneous, do not necessarily resemble actions usually considered as free, or as ones we are held responsible for. Those tend to be slow, based on long deliberations that precede the act itself. Their usefulness in the experimental domain, usually interested in simplification, is questionable, but the same can be said about the actions chosen by Libet. On the other point, the nature of the conscious decision, surprisingly little is said. The readiness potential is throughout analyzed, and the decision seems to be taken to just arrive, almost from nowhere, and is expected to have an impact on the physical processes of the action, or initiate it. Also this assumption seems rather unlikely.

Determinism/indeterminism problem In Libet's approach.

Since the discussions on the compatibility of free will with the deterministic nature of the universe have dominated the philosophical views on free will, the issue also appears in connection to neuroscientific experiments, although on a much smaller scale. Many neuroscientists (e.g. Haggard, 2011) seem to think that the world, and neuroscience, both in its methods as well as assumptions, is deterministic. Therefore, any account of free will to have a scientific standing, has to be compatible with determinism. On the other hand, Libet considered both: determinism, as well as indeterminism, a belief system rather than a scientific theory. He argues

[…] Must we accept determinism? Is indeterminism a viable option? We should recognize that both of these alternative views (natural law determinism versus nondeterminism) are unproven theories, in other words, unproven in relation to the existence of free will. […]

natural laws were derived from observation of physical objects, not from subjective mental phenomena. The latter cannot be directly observed; they are inner experiences of the individual who has them. There has been no evidence, or even a proposed experimental test design, that definitively or convincingly demonstrates the validity of natural law determinism as the mediator or instrument of free choice or free will. (Libet, 2004, p.152- 153)

Against the common belief of the extension of scientific assumptions, and practices to all areas, Libet thinks that although the natural laws were observed in physical reality, and the brain is a physical entity, it does not necessarily mean that the same approach will work for both. Although indirectly, Libet seems to assume some form of mind to body independence, probably not in terms of substance, but rather in terms of irreducibility of the former to the latter. Libet points out that in other science, the object of their inquiry has always been objectively observable, which cannot be said about the mental, this however makes his lack of inquiry into the nature of the connection between the mental and the physical even more puzzling.

LIbet also rejects one of the problematic consequences of determinism, namely the predictability of the future events, when he argues that

[…] I can say that these subjective phenomena are not predictable by knowledge of neuronal function. This is in contrast to my earlier views as a young scientist, when I believed the validity of determinist materialism. […] There is no guarantee that the phenomenon of awareness and its concomitants will be explainable in terms of presently known physics. (Libet, 2004, p.5)

Therefore, although the brain might function in a deterministic fashion, from that does not follow the mental will be predictable. But what is the connection between the two then?

Libet and the mind-body interactions.

Libet, although distinguishing two potential causes of action, the physical and the mental, does little to explain their mutual relationship. Most of what can be concluded about his view on the matter, stems from the way the experiments are constructed and conducted. The sudden appearance of conscious decision, and lack of any account of connection between the two suggests dualism, a position clearly rejected by neuroscientist (e.g. Desmurget, 2009a, 2009b). Where most of them seem to favor reduction, as we will see with more detail later on, which brings a set of problems on its own, Libet seems to go more in the direction of nonreductive physicalism (where the mental, although not independent from the physical, cannot be reduced to it either) or emergence of the mental from the physical, when he states:

In fact, conscious mental phenomena are not reducible to or explicable by knowledge of nerve cell activities. You could look into the brain and see nerve cell interconnections and neural messages popping about in immense profusion. But you would not observe any conscious mental subjective phenomena. Only a report by the individual who is experiencing such phenomena could tell you about them. (Libet, 2004, p.5)

He emphasizes the subjective nature of the mental and explicitly states that mental events cannot be explained fully by the physical brain processes. Unfortunately, Libet does not clarify this fully, sometimes stating that

[…] these physical events are following the natural laws at the macro level. However, that does not exclude the possibility that physical events are susceptible to an external 'mental force' at the micro level, in a way that would not be observable or detectable (Libet, 2004, p.154)

His approach could be understood as a form of dualism, where mental force is somehow different than the physical brain processes. This transpires from the construction of the experiment, and is not ever completely clarified. Under the assumption of dualism, the conscious mental decision would be of a different substance, but under the assumption of nonreductive physicalism, which is much more likely given the setting, the mental seems to appear from nowhere. Without the investigation into the links between the mental and the physical, it is hard to determine what the RP actually stands for, as it might be connected to the movement but just as well can be seen as some form of underlying physical activity connected to the decision itself.

Another strong possibility is that Libet sees the mental as, in some sense, emergent from the physical, where, in the abstract of his article with Haggard (2001), they state:

The problem of free will lies at the heart of modern scientific studies of consciousness. An influential series of experiments by Libet has suggested that conscious intentions arise as a result of brain activity. This contrasts with traditional concepts of free will in which the mind controls the body. (Haggard and Libet, 2001, p.47)

Describing the mental as arising from the physical can be interpreted either as some form of emergence of the former from the latter, or a form of realization and supervenience. In this dissertation, as I will also discuss in the next chapters, I will argue that the latter interpretation is the most likely one.

Thinking of the mind as the control unit, is again unclear, as it can be interpreted as dualistic, where the different mind-substance has the power of the physical brain-substance, and again, this is only applicable to a hand full of philosophical theories and not to the whole field.

Short summary of Libet's experiments.

Libet's experiment aimed to investigate the causes of our overt actions, through recordings done using EEG, on participants asked to perform a simple task. The experiments and their results, as seen in the cases where no pre-planning was reported, are illustrated on the figure below.

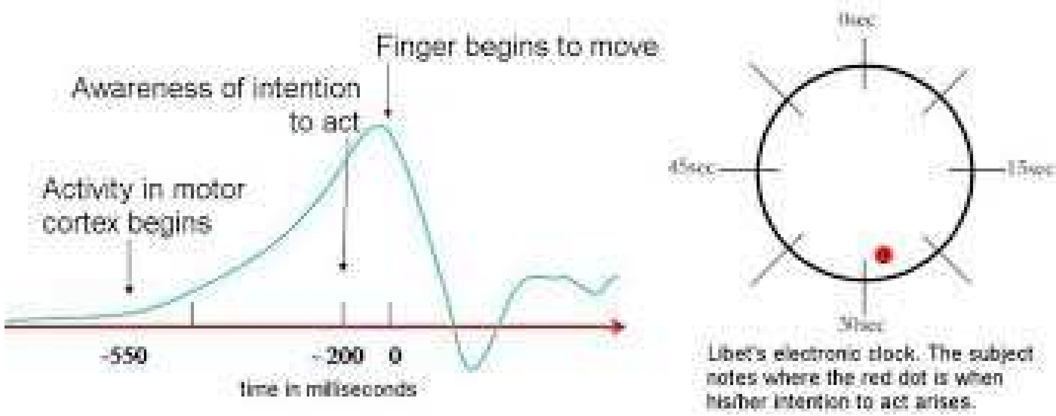


Figure 4. Libet's experiment.[4]

Since its original run, the experiment has been modified and repeated a number of times. The results of those experiments, along with critiques of Libet's original results will be presented in the proceeding chapter.

Libet's legacy: responses to the experiments.

The groundbreaking results of Libet's experiment triggered a broad discussing both in philosophy and in neuroscience, questioning everything from their methods, through results up to the conclusions. The conclusions, and their validity, are still disputed today, both from the philosophical and from the experimental perspective, and the problems range from the type of action that were chosen, to what the readiness potential really means.

Are the actions chosen by Libet really voluntary?

One of the possible sources of problems with Libet's experiments identified comes from the type of actions chosen for the experiment. Can they be called paradigmatic, or a good reflection of our intuitive understanding of what free will is?

[4] Figure from http://www.world-of-lucid-dreaming.com/is-free-will-an-illusion.html?fb_comment_id=10151532036550397_10154386360190397

Some think that Libet's actions can be considered a rather degenerate example of free will (Roskies, 2011), as they lack any rational motivation to conduct them. Such conclusion would suggest reasons are necessary for free action, and potentially can be seen as causal in the action production. This in turn comes with a set of its own problems, which will be discussed further in the next chapters of this dissertation.

Others (e.g. Bayne, 2011) argue that, although actions chosen by Libet are not necessarily automatic, and fall under the voluntary range, the latter should be further divided into *disinterested* and *deliberative agency* (Bayne, 2011), and Libet's actions, although voluntary are of the disinterested rather than deliberative type. In Bayne's approach the willed and automatic are not two opposite types of action, but rather two sides of a continuum, with the two additional subsets of action, being on the voluntary end of it. From the two, only the latter, which requires engagement of reasons and the commitment of the agent to the choice they are making, could be considered a true exercise of free will, and we only exercise deliberative agency when the alternatives we have to choose from matter to us. The actions used by Libet are clearly not of that type. Again the, possibly causal, role played by reasons is emphasized in the type of actions usually considered free.

In Libet's experiments the agents did recognize themselves as the authors of the action, but the exact decision about the moment in which the action was executed, was of little importance to them. As we remember from the previous chapter, the participants were asked to choose the moment in which the action was executed. Each moment was exactly the same as any other one, therefore the choice, although it had to be made, made little difference to the subjects. Bayne then argues that those actions, although voluntary and a result of our intention or decision, are not what we usually identify as an exercise of free will.

Similar arguments are made by Kane (1996), where, e.g. an agent is supposed to choose a breakfast item from two equally interesting possibilities. Nothing hinges on choosing either of the options, and although, in principle, a choice between the two could be described as free, the agent is indifferent to choosing either one of the two. For some philosophers free will can only be truly exercised in situations where the choices made by the agents are somehow significant, where there are consequences, e.g. moral or practical, to the decisions taken by the agent (e.g. Kane, 1989)

The choice of actions was also criticized by Habermas (2008), who argues that for an action to be an exercise of free will, it needs to happen in a specific context, rather than in a laboratory setting. Cultural and societal norms matter in assessing if the movement executed was intentional, or not, and the same movement conducted in a lab has a different meaning that the one made by a cyclist on a road, who is signaling his intention to turn. Free action is only one that is a result of reasons and rational explanation.

The common thread connecting the positions discussed here, ones that question what type of actions are best thought to reflect our intuitions on free will, is the importance given to reasons for action. Here the connection between this discussion and Frankfurt's redefined principle seems clear: the actions can be classified as free when they are caused by, or accompanied by, reasons. In Frankfurt's case the actual causality of reasons does not seem to be necessary, since the action is a result of the previous states of the universe in combination with the laws of universe, so it does not call for the causal role of reasons, but they however have to be present in the action generation processes. In case of criticism of Libet's action, the causality of reasons in generating the action seems to be taken as paramount, and necessary for free action.

What is the origin of our actions? Is motor cortex the appropriate region to look for them?

Libet's experiments focused on the motor regions of the brain, assuming that is where the overt, motor action begins, but also this might have not been the best choice. Soon et al. (2008), who conducted modified versions of Libet's original experiment, argue that they can identify at least two essential problems with Libet's approach: (1) the lack of actual choice between alternatives, and (2) the region of the brain from which the recordings were done.

Soon and colleagues argue that the participants, in the original experiment, do not really choose between different options, like different types of movements, but are only required to make one decision: at which moment do they execute the movement. This leaves open the possibility that the processes investigated do not reflect the decision to move, conscious or unconscious, but rather an unspecified preparatory activity, that is probably linked to the movement, but is not necessarily identical with the decision processes. In their version of the experiment they changed the action to reflect the requirement of different possible outcomes (the participants here had to choose between different letters).

Soon et al. (2008), besides changing the type of action performed, also extended the recording areas, beyond SMA (supplementary motor area), reassessing the time delay between the onset of readiness potential, and the decision the participants were making. They have simultaneously explored if any brain activity can be found to exactly predict the outcome of the choice. Their assumptions seem to be that, although SMA is without a doubt involved in the process of action generation, it is questionable that it is the actual point of origin of the decision to act, conscious or unconscious.

In Soon and colleague's version the subjects were asked to:

[…] relax, while fixating on the center of the screen where a stream of letters was presented. At some point, when they felt the urge to do so, they were to freely decide between one of two buttons, operated by the left and right index fingers, and press it immediately. In parallel, they should remember the letter presented when their motor decision was consciously made. (Soon et al., 2008, p.1[5])

In order to avoid the, already mentioned, issue in the original experiment, where the participants only chose a moment of a movement, and did not performed a choice between different options, Soon et al. asked the participants to choose between two different buttons instead. The free choice of moment is preserved, but the objects of choice are added, therefore both *when* decision and *what* decision are being taken by the participant.

Unlike Libet's, Soon's experiment was conducted using not EEG, but rather functional magnetic resonance imaging (fMRI). After each run, when the participants have pressed the button at the moment of their choosing, they were shown four letters and asked which one of them was visible on the screen at the moment of their decision. Only then the new run was initiated.

[5] Quote is taken from online publication, therefore the page numbers might differ from the standard publication.

The button was pressed, on average 21,6 sec after the onset of each trial, the left and right button were chosen more or less equally frequently and the intention to move preceded the movement by more or less 1000 msec.

The experiments found that the decisions made by the participants of the experiment.t regarding their motor actions, were encoded not only by the primary motor cortex and supplementary motor area, but additionally the lateralization of the participant's decision was encoded in fronto-polar cortex and parietal cortex. The predictive information was identified already 7 sec before the subject's conscious decision, and authors hypothesize that the information could be present up to 10 sec before the movement. The timing of the decision could be decoded up to 5 sec before it, but only from two areas: pre-SMA and SMA. Therefore, the earliest time from which the decision of the subject could be predicted, was extended significantly, both in terms of timing as well as the brain regions that are involved. The authors conclude that the brain activity is

[…]not an unspecific preparation of response. Instead, it specifically encodes how a subject is going to decide. Thus, the SMA is presumably not the ultimate cortical decision stage where the conscious intention is instantiated as has been previously suggested. (Soon et al., 2008, p.3)

Again, do those result lead to a conclusive answer regarding the possibility of free will? Some of the possibility seems to hinge on the predictive part of the experiment, as it seems that if we are able to predict the movement from only the brain activation, free will drifts further from the realm of possibility. As noted by Mele (2014) the accuracy of the tests conducted by Soon et al., when it comes too predicting the outcome of the action, is only around 60%, not much better than the chance-level. The prediction is of course relevant because if we can advance predict what the participant of the experiment will do, it gives us further proof that the conscious intention is not the cause of the action. However, based on this low percentage it is plausible to think that the decision was not in fact taken in that specific moment of time. Rather the prediction rate points to the existence of some form of unconscious bias towards left or right side, since the experiment focused also on the lateralization of the movement and not only on the movement itself. Therefore, the predicted activity might just show that the participant is slightly more likely to choose one side over the other one, which marks a difference between a slight inclination to do something and the intention or decision to do something. The former, although potentially having a causal impact on our final action, is not the same as the intention and decision which we consider an actual cause of the action. What would happen if the prediction percentage increases? It can be also argued that it would just show the connection between the mental and the physical, which in turn is only questioned under the assumption of dualism. In the article by Fried et al. (2011), the authors have recorded the activity of 1019 neurons during the execution of a Libet-like task of executing a self-initiated movement. They have found that there were able to predict an impending decision to move using 256 SMA neurons with 80% accuracy, which is much higher than the one reported by Soon et al. Moreover, they were also able to predict the moment of the decision to move, with the precision of few hundred milliseconds. Yet, again, we can ask if this really tells us something specific about having free will or not, or are the conclusions just contingent on the specific assumptions of the experiment: firstly, on the type of actions chosen, and secondly on the nature of connection between mental and physical. Since if we agree that a free decision is a result of a long deliberation, and that possibly intermediate decisions are taken in the course of it, that we automatize as much as possible (which would make sense given that conscious decision seem costly), and that mental is not a separate substance that arrives from nowhere, those results do not seem as solid. Here I want to argue that, without investigating further the relationship between mental and the physical, the results tell us surprisingly little, and they can be then interpreted in two ways: either as a proof that free will does not exist, or that there is a connection between the mental and the physical, and there might be a specific brain region, connected to the act of choosing or deciding. Said region might not be the motor cortex, as in Libet's experiments. Moreover the results, as the ones gathered by Soon et al., might suggest that the way decision to act is seen on Libet's account is incorrect, and, as we have seen in the previous sections, there are reasons to think that the actions chosen by Libet, and by Soon et al., are not a good example of a free action.

More proof that the action does not in fact originate from the motor regions of our brain, came from studies on patients with brain damage. In a study by Sirigu et al. (2004) a version of Libet's task in conducted on patients suffering from parietal lesions. Unlike healthy participants in Libet-style experiments, or patients that suffered damage to cerebellum, as a result of a stroke, patients with parietal lesions were unable to identify the moment when they have become aware of their intention to move, while still being able to report the moment of movement initiation. In other words, the patients were only able to identify the will to move as the movement was starting. This led the authors to conclude that the planning of a movement, a predictive internal model of that movement is created in the parietal cortex.

Previous research suggests that while cerebellum is linked to the automatic, rather than voluntary, actions (Wolpert et al. 1995), others have already suggested that the parietal cortex might be involved in the awareness of movement and intention (Desumrget and Grafton, 2000; Blakemore and Sirigu, 2003). The current model (Sirigu et al., 2004) suggests that awareness is a part of the cortico-cortical sensimotor processing loop, and parietal cortex might be the neural correlate of motor awareness. In other words the intention to move might not originate in the motor cortex, as suggested by Libet, but rather from the parietal regions of the brain. However, Lafague and Duffau (2008) were unable to repeat the results obtained by Sirigu et al. (2004) using patients that had their inferior parietal lobule resected. The authors, based on their result, conclude that it is strongly suggested that this specific part of the brain is involved in awareness and voluntary action.

All of the research discussed above questions if motor cortex is in fact the best area to investigate, while looking for the origin of conscious decision to act. If so, Libet's paradigm, and other methods of investigating decision making, should have the area of recording extended, to include other regions of the brain than just a motor cortex, and potentially record at different timings than Libet.

What is RP the mark of?

Further experiments on the relationship between readiness potential, conscious decision, action and the possibility of free will bring other mixed results. Given the way Libet's experiments were designed, the readiness potential was always accompanied by the movement. As stated before, it was necessary for practical reasons: the brain activation had to be backtracked from the moment of the movement, which was the objective marker for the scientist. But what happens when there is no movement? Is the RP still present?

Alexander et al. (2016) conducted a version of the experiment with those exact questions in mind, and the results they obtained pointed to the existence of the brain activation also in cases where no movement is present. These issue is of great significance, since most of the arguments against free will in neuroscience, relies on the assumption that RP is somehow associated with unconscious processes of movement initiation. If that is not the case, it is hard to substantiate Libet's conclusions that the brain clearly decides before the conscious decision is made. In their experiments they asked the participants to choose one of the four letters presented to them (similarly to the paradigm used by Soon et al.). Subsequently participants were asked to, in half of the trails press a button at as soon as they have decided, while in the other half they were not required to make any movement at all. The first type was then used as a blueprint, and the two different types of responses were time-locked together, to the subject's time of decision that he noted and then reported back. This allowed for the comparison between the two different runs of the experiment: with and without the overt movement.

The subjects were again monitoring a clock, and a new run started whenever they pressed a button in each of the trials. During the duration of the trial a random letter appeared in each of the clock's quadrants. Each of the letters was presented to the participant for 150 ms, and after that was replaced by a different letter. Once the clock reached the position identified as 3 on the standard clock, the letters stopped changing, however their color kept on switching with the same speed (Alexandre, et al., 2016, p.40). The subjects could choose between the four letters presented to them, between the 3 o'clock when the letters stopped changing, and 12 o'clock. They were also, just like in Libet's original experiment, asked to remember the position of the clock at the moment of their choosing. The brain processes were again recorded using EEG with electrodes, placed in 64 different scalp locations.

In both variants readiness potential was recorded, and the two types did not differ significantly, suggesting no difference between the movement and no-movement conditions, as far as the brain activation goes. This led authors to conclude that the process

(RP) is not related specifically to the motor preparation processes, as the strength of the RP was the same in the trials with no movement. It might be related somehow to the decision or is an anticipatory process of some kind instead, and might not be related to the preconscious motor planning after all.

A different approach was taken by Schlegel et al. (2015), who, instead of investigating the existence of readiness potential under no-movement conditions, decide to analyze brain activations in cases where the actions are not voluntary. To do so, their experimental setup included movements made under suggestion, using, in the first phase, hypnotic induction, and then post-hypnotic suggestion. Hypnotic induction is

[…] a process of inducting a trance-like state via guided imagery (Schlegel et al., 2015, p.197), while the post-hypnotic suggestion involved

[…] an instruction given to a hypnotized person during hypnosis that is to be followed after waking from the hypnotic state (Schlegel et al., 2015, p.197).

The whole experiment consisted of five phases: *[…] EEG preparation, first hypnotic induction, first task phase (with post-hypnotic suggestion), second hypnotic induction, and second task phase (without post-hypnotic suggestion).* (Schlegel et al., 2015, p.197).

For the first task the participants were watching silent nature videos, and they were supposed to squeeze a stress ball, one of which they held in each hand, when a red arrow appeared on the screen. Depending on where the arrow appeared, the participants were supposed to either squeeze the ball with the right or their left hand. Hypnosis was used to elicit movement without the sense of conscious will. No difference in the activation of the RP was recorded in comparison to the voluntary movement, which led authors to conclude that the process itself can be unrelated to both the conscious intention to move and the conscious decision itself.

The results obtained by these experiments put into question Libet's conclusion on the close link between the readiness potential and the initiation of action. The fact the RP is present both in cases where the movement is absent, as well as ones where the decision is absent, suggests a different role played by the brain activation, and the lack of connection between it, the conscious decision, and the action.

Readiness potential or the lateralized readiness potential?

The role of readiness potential in the action initiation was also questioned by Patrick Haggard, who conducted a variety of modified Libet's experiments. Similarly to the latter, he argues that the neural spikes preceded the conscious decision to act, and therefore the latter is not the initiator of the movement. Unlike Libet, he argues that we do not have free will and seems to reject the free won't advocated by his predecessor (therefore giving no causal role to the conscious decision whatsoever).

In one of his more influential papers, Haggard, along with Eimer (1999), attempts to clear out some of the issues regarding the relationship between the readiness potential and the awareness of movement. They argue that the content of the events described by Libet, both the unconscious readiness potential, as well as the conscious decision, is unspecific. Therefore it can be a mark of a number of things, from a very general state, e.g. arousal, to very specific ones, e.g. left finger flexion (Haggard and Eimer, 1999, p.128). Moreover, since Libet, there has been a shift of paradigm and more frequently the lateralized readiness potential rather than just readiness potential is used as a mark for specific motor preparations.

Haggard and Eimer's experiment had three main aims: (1) to compare M (movement) and W (decision) judgements in the fully spontaneous and fixed movement conditions, (2) measure the LRP associated with movement and study its relation to the W judgement, and (3) testing Libet's hypothesis regarding the relationship between the W judgement and brain potential, where a variation in the time of the former was associated with a variation in the latter. The last hypothesis is most interesting from the point of view expressed in this dissertation, as it has the potential to, to some extent, lay the foundations of investigating the connection between the mental and the physical. Assuming covariance of the two, some form of supervenience relationship, which is generally accepted in nonreductive physicalism position, might be suggested. Secondly same hypothesis provides an initial insight to the argument by Mele (2009), who argued that the consistent spikes in brain activity might be in fact a preparatory process supporting an intention forming.

Haggard and Eimer (1999; discussed also in Haggard and Libet 2001) conducted two separate versions of Libet's experiment: one using his original method, and a modified one, where the subjects had to additionally choose which hand to conduct the movement with. In the latter experiment, lateralized readiness potential related to the choice between two hands was recorded.

The general conclusions of the experiments suggest that the judgements (both M and W), which Libet assumes are general, are in fact specific and tied to the movement, therefore do not reflect any higher level processing. There was also no difference between the two judgements in the two different types of movement (free, and fixed movement conditions). It is hypothesized that, if they were reflective of higher level abstract planning, there would be a difference between the two conditions, as the free movement would, presumably, require longer preparation. According to the authors' initial hypothesis, if RP is connected to the intention, it should appear earlier, in cases of actions resulting from a distal intention, and later, for the ones that were a result of proximal intentions. Yet no such relation was found in either one of the versions of the experiment, which led them to the conclusion that the RP, as recorded by Libet, could not be identified with the intention. Interestingly, the results obtained from the lateralized readiness potential were more promising, since for the trials where the intention occurred early, so did the lateralization of RP. This suggests some form of connection between the two.

Schelgel et al. (2013) attempted to repeat the results obtained by Haggard and Eimer using a larger group of participants and several variations of that method, but the trial proved to be unsuccessful. Also, no correlation was found between the LRP (lateralized readiness potential) and the W judgement (conscious decision) making also this process potentially independent from the decision itself and not causal in the process of its forming, through the forming of an intention.

Trevena and Miller (2002) showed that LRP may in fact follow, rather than precede, W time. In their later work (Trevena and Miller, 2010) they show that RPs and LRPs, occurring before decisions to move, were not significantly different from these signals preceding decisions not to move. Thus, one possibility is that the RP and LRP are related to general readiness to act, rather than to the specific decision to act now (Pockett and Purdy, 2010). And there appears to be no clear-cut evidence that either RP or LRP are neural markers of intention onset, or that they are part of the causal chain leading to intention.

What does the W judgement stand for?

Both in the original experiments as in the modified versions, the W judgement is unequivocally recognized as a moment of when the conscious decision appears (the wanting moment). Peter Ulric Tse, whose theory will be discussed more in detail further in the dissertation, attempted to determine what W judgment actually stands for, and he identified several possibilities: (1) conscious feeling of being about to move (2) intending to move (3) having an urge or desire to move (4) feeling that an imminent motion is agentically authored by oneself (5) prospectively causal content of the subsequent motion (6) only retrospectively causal of motion (7) conscious correlate of a form of efference copy that marks motion execution as self-caused, as opposed to be caused by external factors (Tse, 2013). Given Libet's paradigm, all of those interpretations are consistent, and we have no means to rule any of them out.

This ambiguity as to what the W judgement might mean does not necessarily give us enough reason to question the validity of the measurements. Yet, the exact role of the W judgement, depending on which of the readings is correct, might play a different role in the movement execution.

Others found that the time of conscious decision might be inferred rather than felt. In a recent study by Rigoni et al. (2010), the authors conducted a modified version of Libet's experiment where participants were asked to press a button and report the time of conscious decision. Additionally, they have provided delayed auditory feedbacks which were there to signify the action, although they appeared after the action itself. They have found that:

[…] auditory feedbacks evoked a negative component in the 250-300 time range, namely action-effect negativity […] [which amplitude] *was sensitive to the delay of the auditory feedback, with a larger amplitude for more delay feedbacks. Furthermore, changes in the*

[action-effect negativity] were also associated with changes in the reported W. (Rigoni et al., 2010, p.1)

Therefore, they found clear evidence that the auditory feedbacks influenced the time of conscious decision, which supported the initial assumption of the experiment. This means that the participant could have inferred rather than felt the decision to move, as the feedbacks had an impact on the identified timing of the decision. On the basis of the obtained results, they concluded that not only we infer when we have decided to act from what happens after, but also that the subjective experience itself is influenced by a system that tracks if there are any mismatches between the actual action and the action prediction.

A similar experiment was conducted by Banks and Isham (2009), where a beep was presented to the participants after the movement was already executed. They have observed that the time of decision moved linearly in time along with the beep, which means that the auditory signal was again found to have an impact on the reported time of making a decision. In a second version of the experiment, the participants were looking at their hand conducting the movement with a slight delay, which also influence the reported time of decision. Again the decision was reported to have occurred slightly later than on the trials without the observed delay of the hand movement. This also lead the authors to conclude that the reported time of the decision is inferred from what happens later, the movement and what occurs right after the movement. Other experiments (e.g. Lau et al., 2007) used transcranial magnetic stimulation, applied on the preSMA, and were able to shift the perceived intention to act backwards in time, while perceived time of action execution was moved forwards.

Focusing on the clock rather than on the conscious intention? Issues stemming from using EEG.

Claims made by Libet depend largely on the readings of EEG, and his assumption that the activity recorded is in fact linked to the movement preparation. Such assumptions are not exclusive to Libet but were also made by those who conducted multiple variations of his experiments (e.g. Haggard and Eimer, 1999; Trevena and Miller, 2002) or revisiting the notion of readiness potential itself (e.g. Shibasaki and Hallett, 2006).

The activity recorded in all experiments, essential to its finding, is a negative shift in the EEG, recorded at the C_z electrode. However, although it is generally accepted that these types of activity happen before voluntary movements, it is logically impossible to show that this shift actually arises in conjunction with the movement preparation (Miller et al., 2011). This can be considered a crucial problem, since if the consistent appearance of RP is not in fact connected with the initiation of the movement, but rather with some other brain process, it can no longer support the claim that movements begin unconsciously.

One of the possible interpretations connects the negative shift with the participant's monitoring of the clock while undergoing the experiment, along with monitoring the appearance of the intention to move. In the paper by Miller et al. (2011) two different experiments were conducted, and both consisted of two different runs. In the first variation of the first experiment the participants were asked to make a hand movement while watching a rotating clock, at the moment of their choosing, just like in Libet's experiment, while in the second version the participants were asked to simply move their hand at whatever moment they chose, without having to observe the clock. The results were conclusive:

In sum, the requirement to monitor the clock and report the position of a moving spot at the time of the spontaneous decision to make the key-press response altered the average EEG amplitude preceding that decision. The effect show in the response-locked averages suggests that the pre-movement negativity at Cz reported by Libet et al. (1983) may have resulted from the processes associated with monitoring the clock rather than from unconscious preparation for movement. In addition, the effect shown in the stimulus-locked averages demonstrates that—not surprisingly—EEG activity at the start of a trial was also affected by the onset of the clock, possibly because of the obvious differences in visual stimulation between clock and no-clock trials. (Miller et al. 2011, p.105)

In other words, the results of the first experiment suggest that, rather than the initiation of the action, the RP is connected to monitoring of the clock by the participant. Much of the activation recorded was due to the clock being present and had little to do with the initiation of the movement.

In the second experiment the assumption was that the brain activity, if in fact connected with observing the clock and not with the preparation to move, would be present also in conditions where no movement was required. To test that an experiment has been conducted where participants, rather than moving a hand, were asked to make a perceptual judgement. Again two trials were run, one with and one without the clock. Also here there differences between trials with and without the clock were detected, however the statistical significance of the results was smaller than in the first experiment. The results clearly challenge Libet's conclusions about the unconscious initiation of the voluntary movement, given the methods used.

Does belief or lack of belief in free will changes the response of the brain?

In one of the previous chapters I have discussed the influence of folk belief in free will on the behavior and attitudes of the individuals. As was shown there, reading a text falsifying the existence of free will had a direct impact on individual's behavior. Similar can be also observed in the realm of physical brain activations.

Said influence was investigated by Rigoni et al. (2011), who conducted an experiment where the participants were yet again asked to perform a Libet-style experiment. In this case the authors were tracking both the late RP (as seen in Libet's experiments) as well as the early RP, which appears around 2000 ms before movement execution, as observed by Kornhuber and Deecke (1965). Their hypothesis checked was that the disbelief in free will would lead to a change of amplitude in the early RP. Similarly to the experiments conducted by Vohs and Schooler (2008) the participants were divided into two groups, each of them given a text to read.

The first group, no-free will, was given a text claiming that scientists have conclusively shown that free will is just an illusion, while the other group red an article about consciousness that did not mention free will. After reading the text, the subjects performed the Libet experiment (without major variations) and subsequently, to assess whether they beliefs were in fact influence by the reading, they filled the Free Will and Determinism Scale (in this case the authors used the scale from Rakos et al., 2008).

The scale itself already revealed statistically relevant differences between the two groups, which led authors to conclude that the reading was successful in influencing the subject's beliefs about free will. The authors also checked if the manipulation of their belief in free will had any impact on the reported time of the W judgement, but no such correlation was discovered. The readiness potential, on the other hand, was affected. The study showed that a decreased belief in free will impacted the early RP amplitude, which was significantly lower when compared to the control group. Even more interestingly, the results of the experiment seem to bind the belief in free will with the early motor preparation rather than the motor execution, as investigated by Libet (late RP), since the latter showed no difference between the no free will and the control group. How this affect intention and intentional effort is a still open question as no studies (that I know of) have conducted experiments on this problem.

What the results of this experiment seem to show is that our belief system might have a more fundamental and reaching effect on the decision making process than previously thought. Although the relation with intention forming is unknown, its impact on the unconscious processes seems evident.

Daniel Wegner and the illusion of conscious will.

A completely different angle in investigating the possible existence of free will is taken by Daniel Wegner, who, based on the study of automatisms, concluded that conscious free will is simply an illusion, or rather it does nothing in terms of causing an action.

Wegner argues that two, separate, neuronal pathways, exist one for the actual action production, and the other one supporting the conscious will. The action is produced, just as the conscious will, however the latter does not play a causal role in producing the former.

According to Wegner:

[..] people experience conscious will when they interpret their own thought as the cause of their action. (Wegner 2002, p.64 after Wegner and Wheatley 1999),

Through the experience of will, we can observe a dichotomy between real and apparent cause, and the conscious will is an epiphenomenon providing an emotion of authorship, in order to build a meaningful (retrospective) attribution of our actions. But, as far as the actual overt action is concerned, it has no causal impact.

Wegner recognized two different types of will- empirical and phenomenological. The former is defined as the causality of individual's thoughts, an object of a mechanistic explanation that we are able to analyze with scientific tools, determining how it can influence the individual's behavior. The latter, on the other hand, is the person's reported experience of will, an object of a "mentalistic" explanation limited to understanding other minds and not applicable to physical universe. In Wegner's argument about the phenomenological will, the individual's conscious will, is non causal, and the apparent mental causation appears when people, while experiencing the conscious will, interpret those thoughts as the causes of their actions, when in fact the genuine causes lie elsewhere. The only will that matters is the empirical one, responsible for all the activities of the body.

As for the two neuronal pathways, the action is caused by its unconscious neuronal basis, while the conscious will is caused by a neuronal basis of though, which is also unconscious. On this picture, the thoughts and the actions are in fact a result of completely unconscious brain activations. In other words, for every action, e.g. a movement of a hand as in Libet's experiments, there is a neuronal process causing the action. Aside from this process, there is what we perceive as a conscious decision to execute the action, in this case to move, caused by a different process than the one leading to the overt action. Wegner claims that although we perceive the thought as causal in action generation, it in fact has no causal impact on the process of action production, and is completely epiphenomenal in this context. Wegner allows for there to be some form of interaction between the two causes: (1) of the action, and (2) of the though, however on his account it remains unclear what that relationship could be, and in which direction.

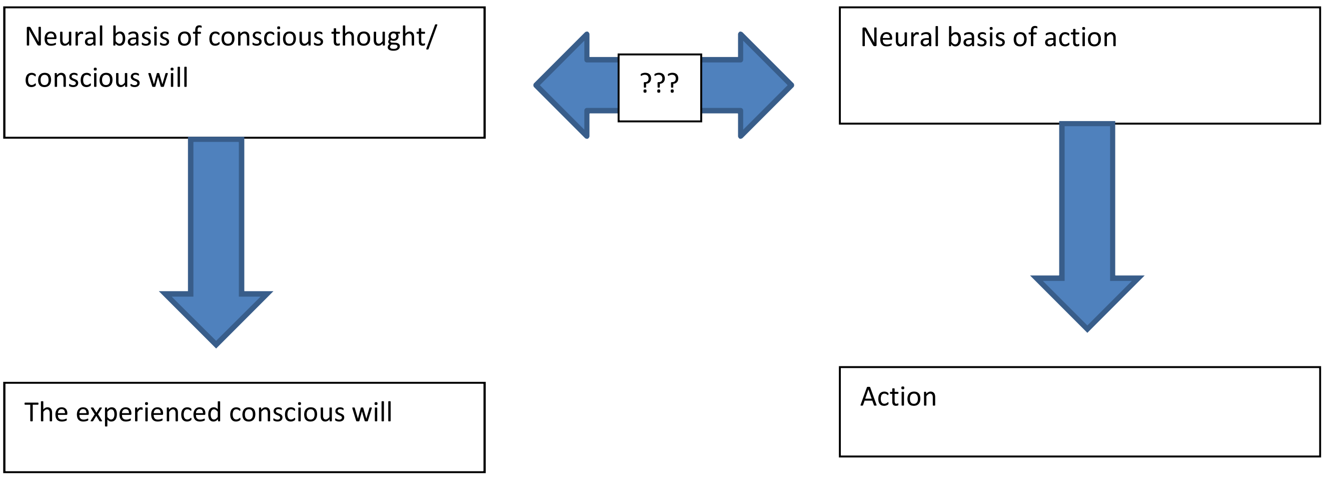


Figure 5. Simplified assumptions of Wegner's model

As we can see on the figure above, the connection between the two neural bases could be, in principle, interpreted as running in either, or both directions, meaning that the neuronal basis of thought could be the cause of the neuronal basis of action, and conversely.

Wegner's conclusions rely heavy on a specific interpretation of causation, where it is identified with an event rather than a property, which means that our experience of will cannot in itself be causal (Andersen, 2006), and people's experience of the conscious will happens regardless of any actual connection between the will and the action itself. The core of the argument in *The Illusion of Conscious Will* is the assumption that the will itself cannot be causal in the process of action initiation. But is the will itself ever causal? If we look at the way Libet framed the problem, the expected possible cause is a conscious decision, and not our experience of it, or our conscious will in general. Therefore something very specific, as opposed to the general picture Wegner seems to be painting. For him, what should be causal is then something close to the experience of willing, which, as a concept, is not clear at all.

I consider experiencing our own conscious will, or our own agenthood seems to be similar to experiencing thirst. It plays a role in our decision to go into the kitchen and pour ourselves a glass of water, however, what we consider truly causal here is the decision, not the experience of thirst. In a way the experience of thirst is a contributing factor in the resulting action, as it is the basis for us deliberating on a solution, forming an intention and in the end going to the kitchen and getting a glass of water, but by itself is not necessarily seen as the sole cause.

Conclusions.

In this chapter I have presented the experimental approach to the study of free will, focusing mainly on Libet's experiment. The experiment itself has become a standard when investigating the existence of free will, although its initial aim was to investigate the causes of our actions rather than, directly, the problem of free will. In the experiment the participants were asked to perform a simple movement, at the time of their choosing. Throughout the duration of the experiments the participants were connected to the EEG machine that was recording their brain activity prior to the onset of the movement. Two potential causes of the behavior were identified, the unconscious brain activation (the readiness potential) and the conscious decision. The experiment and its results are presented on the illustration below:

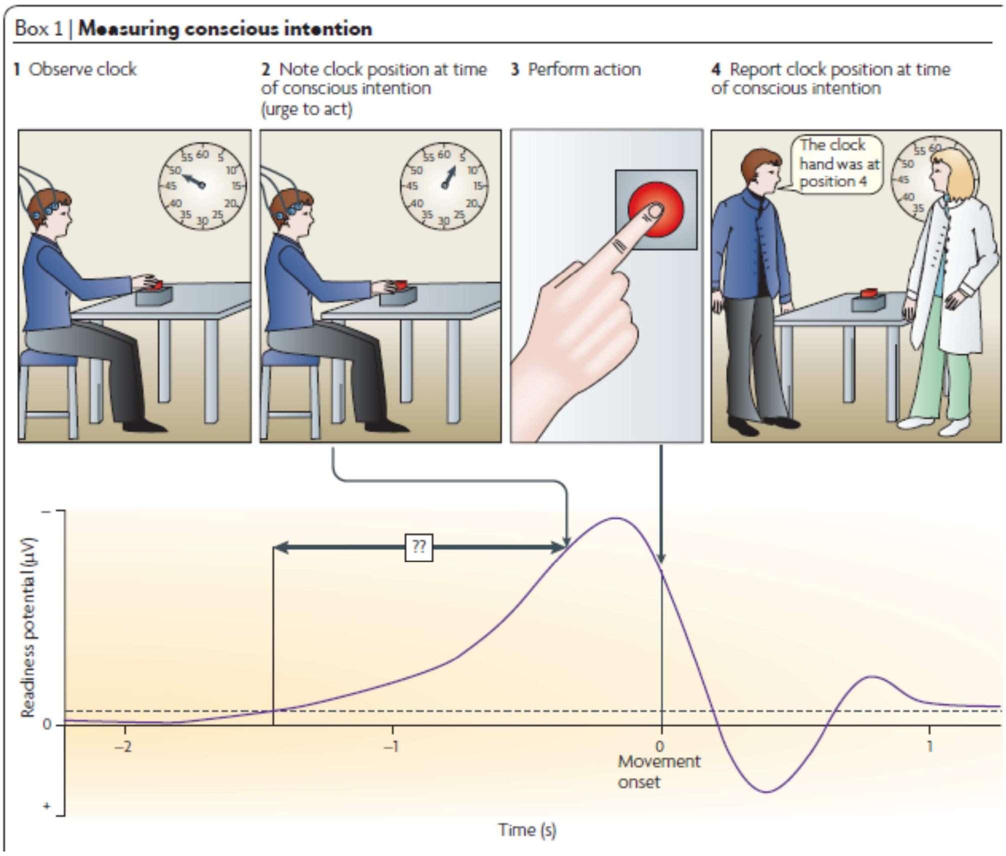


Figure 6. Libet's experiment (Haggard, 2008, p.935)

The results of the experiment suggested that, since it is recorded before the reported time of the conscious decision, it is the brain activation, or the readiness potential, that is the cause of the action. These results led Libet to conclude that conscious decision does not play a role in the action initiation, as would be expected for what is normally understood as free action. However the causal power of the decision is preserved in what Libet calls the veto power, where the decision is the control unit, and can stop the action already underway.

I have then discussed a number of criticisms in connection to Libet's experiment, both from philosophical as well as experimental perspective. One of the arguments often raised in response to Libet's, and the variations of the experiment that followed questions whether the mere fact of the existence of the physical brain activation closely connected to the overt action and proceeding the mental event is all that problematic. The challenge raised by neuroscientists here seems to mainly hit, and is often raised in connection to (e.g. Desmurget et al. 2009a), dualism, a position that is controversial also in philosophy. The conscious decision appears from nowhere, and not brain activations connected to it are investigated, suggesting its, at least partial, independence from the physical realm. Another serious problem stems from the type of action chosen by Libet, which seems to be, although voluntary, still different from what is usually considered a free action. The objection was usually raised in connection to the lack of reasons for the actions used in the experiment.

In the discussion so far, one essential feature of the free action seems to come to the foreground, namely the way the action is caused. The causation lies not only in the neuroscientific discussions, where the causes of action are all that matters, but also in the philosophical ones, where the main threat to free will is causal determinism.

Summary of the discussions so far:

1. Philosophical discussions are dominated by attempts to show free will's compatibility, or lack thereof, with the deterministic nature of the universe. Determinism can be understood as causal determinism, where both the types of causes as well as the way action is caused play an essential role. The discussions usually concern external conditions for the existence of free will (e.g. the nature of universe, the existence of alternative possibilities).

2. The discussions in neuroscience focus on internal conditions of how the actions have to be initiated, or caused, in order to be considered free.
3. The main source of data taken to disprove free will are taken from Libet's and Libet-style experiments, which operate under several, often implicit assumptions, that in turn can have a big impact on the interpretation of the data. One of the core ones seems to be the presumed independence of the mental and the physical realm, as neither the causes, not possible "grounding" of the mental in the physical, are investigated.
4. There are good reasons, both philosophical as well as methodological, not to think that Libet's experiments conclusively disprove the existence of free will.
5. The experiments however emphasize an important, and often overlooked in the philosophical discussions issues, which connects the problem of free will with that of causation.

Where do we go from here, and what is my position:

1. As was already emphasized, an account of free will that has the types of causes of action at heart requires a solid definitions of causation and of the type of the cause that is needed for free will to be possible. These problems will be discussed in the chapters immediately following the current one.
2. What causes the action, then identified as free, seems important both in the philosophical discussions as well as in neuroscientific ones. However both their nature, as well as the connection between such event and the physical brain activation is often not elaborated on.
3. The way the causes of our actions required for the existence of free will are being described, as conscious decision or intention, suggests their mental nature, which I consider a further proof that clarification of both the notion of causation and a cause, as well as establishing what kinds of links between it and the physical basis hold, is necessary.
4. Although I argue that Libet's experiment do not prove anything about free will in general, they point to a very important features of free will, as exercised in action, including its causes and the definition of the action itself, which, as we will see further is not obvious.

5. The three notions: determinism, predictability and causality are discussed together. However, it might be the case that they come with different problems. Determinism is understood as deterministic causality, and lack of other possible causes (no alternative possibilities). Predictability, often discussed in connection to the neuroscientific experiments, makes the action predictable only on the basis of the neural events, brain processes. This understanding of predictability seems to be based on different intuitions than one of Laplace. It does not require any knowledge of the antecedents nor the laws of nature. Therefore, it can be questioned if such predictability really threatens free will at all (one of the possible answers here would be: of course my actions are predictable, they are also predictable for people that know me as a person, without the necessity to scan my brain). Causality, independently, is hardly ever discussed. It is not completely clear what kind of notion of causality is used here (the possibilities will be presented in the proceeding chapters), and, while for the discussions on causal determinism, this might not be necessary, but, when we frame the discussion on the types of causes necessary for free action (as is done in neuroscience), this becomes crucial.

Chapter 4. What is causality?

The previous three chapters focused on the different ways in which we can approach free will. From the classical debates, where it is discussed in connection to the way the world is organized, through the existence of alternative possibilities, to the action initiation approaches found in neuroscience. I have argued that classification of an action as voluntary greatly depends on the way the action is caused: to be considered free it has to be caused “in the right way”. Therefore, aside from the usually discussed problems like the state of the universe (and causal determinism), and the principle of alternate possibilities, other issues can be identified, found on the intersection between causality and free will. The former notion, similarly to the latter, is complex, and lacks a one-fit all definition, with disagreements ranging from what is causality to questioning if it exists at all. Because of those complexities, and the importance of the notion to the problem of free will, it is essential to clearly define what kind of concept of causality is most applicable here, and which one is used in the debates surrounding the question of the causes of our actions.

The first discussions on causality can be traced back to Aristotle, who, both in *Physics* and in *Metaphysics* offered an account of his approach to causality, where four different causes can be found: material, formal, efficient and final). Moreover he linked two ideas together: that of causality with that of necessity (discussed e.g. Stein, 2012). The same necessity with which the effect follows the cause was also emphasized by Spinoza (e.g. *Given a determinate cause the effect follows of necessity, and without its cause, no effect follows.* Spinoza, *Ethics,* Book 1, Axiom III[6]) and Hobbes, but the real revolution was brought about by Hume, who’s views opened a new chapter in the debate, and became a reference point to much of the discussions that followed.

Many contemporary approaches draw from the system developed by Hume, although this is not necessarily the case for all of them. The dominant ones, as relevant to the discussions presented in this dissertation, are: (1) regularity, and here I will focus on the arguments of Hume and Mackie, (2) production accounts, where besides the classic Salmon and Dowe approach, I will also include the mechanistic causality of Craver, and (3)

[6] I used an online version of the book, therefore no page number, nor the year was given

difference-making account, focusing on Woodward's formulation. All three will play a major role both in identifying the main issues for free will from the angle of causality, as well as in the solutions I will propose in this dissertation. Production accounts and difference making accounts differ radically from one another, as they conceive of the meaning of the notion of a cause differently. For the former being a cause means producing the effect somehow, while for the former the cause is what makes a difference for the occurrence of the effect.

The various accounts found in the literature differ not only in the way causation is identified, but also in the questions they ask about causality, most notably 'what is causality' and 'how can causality be identified'. The latter approaches are not necessarily interested in answering if something as causality exist, but rather focus on investigating what is usually defined as a causal link. The latter can just be a comfortable way to grasp world's regularities, while not being a reflection of something specific and general (as causality would be).

I will start this chapter by identifying some of the most basic concepts involved in the discussion on causality, including the various questions that are often asked (e.g. what causality is vs. how it is identified). I will try to frame the problem in a way that is relevant for the free will debate, and emphasize what kind of questions could be asked in this context. I will then proceed to the classical discussions on causality, starting from Hume, through Mackie, Anscombe and Davidson. I will also include an introduction to counterfactual theories of causation, as counterfactual analysis will be paramount in the following chapters, and Mackie's INUS conditions, which will be important for the fundaments of the position I want to begin developing in this dissertation. From there, I will proceed to discussing the more contemporary approaches, including: (1) the production accounts, (2) interventionism and (3) mechanistic analysis, and I will also discuss what the neuroscientists seem to understand as causality (and to this problem there might be no straightforward answer either).

Causality, basic concepts.

What do we usually mean by saying that an event A was a cause of an event B? In our daily life we seem to have no problem recognizing the connections that make up a causal chain of events. We, often correctly, identify causes of specific events, and we even think that at least some of the events are caused by us. This picture, although functional in our daily existence, does not explain why we think of certain events, as connected in that sort of relationship. Moreover, in philosophy, many question if causality, as something 'in the world', exists at all, or is just a comfortable manner of describing events and variables.

When taken to exist, causality is often defined as being some form of necessary relation, where some, exceptionless, connection between two events exists, which means that the two always follow one another, with the cause being the antecedent of the effect. From this very general description, three features of causality can be identified: (1) it is a type of **relation**, (2) there are **laws** that govern it, and (3) there is a **time lapse** between the cause and the event (most likely they are not synchronic but diachronic). The specific answers to those questions are what distinguish the positions from one another.

Since J.S. Mill it has become common to frame the discussions on causality in terms of necessary and sufficient conditions, where the cause is both necessary and sufficient for the effect to occur (although, as we will see later on in Mackie's approach, those can be understood in number of ways, and can play different roles in actual generation of effects). Cause being necessary is usually understood as: whenever the event A appears, event B always follows, without any exceptions. Sufficient conditions, on the other hand, suggest that, although the event A is enough for the appearance of the event B, event B could occur even if the event A would not. Simplifying it we could say that my thirst is sufficient for my drinking a glass of water, since every time I am thirsty I drink a glass of water (without exceptions). But there are situations where, although I am not thirsty, I still drink a glass of water, for instance, because I think staying hydrated is generally good for me. Identification of the necessary and sufficient conditions for an event disjunctively seems to be much less of a challenge than identifying the one necessary and sufficient condition, that would assure, without a fail, that the effect will occur.

The numerous parallels can be seen connecting the notion of causality and the notion of free will, including, among others: (1) the question of determinism, where causal links, if they exist, can be seen as deterministic or not, and (2) the way the question itself is approached. From one angle we can focus on the metaphysical aspects of both causality and free will, asking about the ontological status of either one of them. From another, we can examine the processes (if they are in fact processes) 'in action', where we try to identify what makes a link causal (is it just a mode of description?). And just like in the free will debate, also in this one, there will be some that claim causation does not exist at all.

In the next section of this dissertation I will focus on Hume and the regularity approaches to causation, trying to establish the main assumptions and hypothesis regarding the notion, used by the authors. It is important to note that the main goal of this chapter, rather than being exhaustive, is to explicitly mark the most relevant notions about causation, which will then serve as a basis in: (1) the analysis of the type of causation necessary for free will, and (2) the understanding of causation usually used in the discussion both on free will, as well as on action initiation.

Hume and the regularity approach to causation.

Most significant contributions made by Hume to the debate on causality can be found in the first book of *A Treatise of Human Nature* and in the *Enquiry concerning Human Understanding.* His theory of causation has continued to have an impact on how causality is defined in modern discussions, yet, in spite of being so influential, it remains surprisingly open to various interpretations, with some reading Hume to be a causal realist, others a projectivist, a reductionist and a realist (Garrett, 2009). Most commonly however, Hume's approach is known as the regularity view on causation, where a cause is defined as:

*An object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter (*Hume, 1739[7])

The cause is therefore defined based on a consistent co-occurrence of a presumed cause and a presumed effect, and where all the occurrences of the cause are followed by the events of the type effect. The two then are *constantly placed in like relations of succession and contiguity* (Hume, 1739), where the cause always precedes the effect. This feature, the lapse of time between the cause and the effect, or the diachronic relation of the two, has become a standard. Some synchronic positions have been developed throughout the years (e.g. Blondeau and Ghins, 2013), however the diachronic approach is still the dominant one.

The conjunction between the two, the cause and the effect, is constant and the cause is efficient, which means that if the latter is not the case, the cause of any kind does not exist. This specific feature of Hume's approach was summarized by Anscombe (whose approach I will discuss with more detail in the proceeding part of this chapter) where:

For he [Hume] *himself assumed that necessary connection is an essential part of the idea of the relation of cause and effect, and he sought for its nature. He thought this could not be found in the situations, objects, or events called 'causes' and 'effects', but was to be found in the human mind's being determined, by experience of constant conjunction, to pass from the sensible impression or memory of one term of the relation to the convinced idea of the other. Thus to say that an event was caused was to say that its occurrence was an instance of some exceptionless generalization connecting such an event with such antecedents as it occurred in.* (Anscombe, 1971, p.90)

The human mind is being determined into believing that there is a causal link between two events, because of the constant conjunction of the same cause being followed by the same effect experienced in the past. However, for the relationship to be more than just a constant conjunction of the two, there needs to be an exceptionless law that covers, and governs such relationships. The determination of the mind through that experience of constant conjunction is very strong, and the mere thought about a cause, is automatically followed by the thought about an effect.

[7] Version used is available at: http://www.davidhume.org/texts/thn.html

Hume's theory puts causal laws as the fundament of causality, having causal relations supervene on them (discussed by e.g. Sosa, E., Tooley, M. 1993). The general rules to determine if the relationship that holds between two events is causal, are given by Hume (1793) in 8 points:

(1) The cause and effect must be contiguous in space and time.

The causal relationship is therefore one that holds between two distinct events, separated from one another with a time lapse (diachronic).

(2) The cause must be prior to the effect.

(3) There must be a constant union betwixt the cause and effect. It is chiefly the quality that constitutes the relation.

The relationship holds between the two without exceptions, and it is constant. I understand it to mean that every time the cause appears, the effect follows.

(4) The same cause always produces the same effect, and the same effect never arises but from the same cause. This principle we derive from experience, and is the source of most of our philosophical reasoning. […]

(5) […] where several different objects produce the same effect, it must be by means of some quality which we discover to be common amongst them. […]

(6) […]The difference in the effects of two resembling objects must proceed from that particular in which they differ.

Based on the last two principles, we can then assume that two identical objects, if they are causes of something, they will be the causes of identical things. Conversely, if two effects, although produces by similar objects, differ from one another, their causal origin has to be linked in some way to that difference, rather than the resembling features.

(7) When any object increases or diminishes with the increase or diminuition of its cause, it is to be regarded as a compounded effect, derived from the union of the several different effects which arise from the several different parts of it cause. […]

(8) […] any object, which exists for any time in its full perfection without any effect, is not the sole cause of that effect but requires to be assisted by some other principle, which may forward its influence and operation.

As was mentioned already above, the core of the relationship between cause and effect lies in causal inference, a consequence of the causal relation, and the basis for the discovery of it. Itself, it comes from *constant conjunctions* where one event is always followed by the other. From this constant co-appearance of the two events, we then form the inference of the existing causal connection between the two. This however, is met by, probably, the most famous contribution of Hume, namely the problem of induction.

According to Hume the simple fact we experience the constant conjunction between the cause and the effect, does not provide us with enough material to suggest that we can infer it will always be so (in the future). Using the example of the sun, the fact that the sun rose every day since the beginning of time, does not give us enough data to conclude that this will be the case also tomorrow. The instances we have not yet, or not at all, observed would have to be the same as the ones we have observed, which cannot be supported by the data we currently have. As is well known, Hume argues that the principle of induction cannot be proved either deductively, or inductively.

The full structure of the argument on causation made by Hume can be then summarized as follows (from Mackie, 1974, p.10):

(a) Causation in the objects, so far as we know, is only regular succession.

(b) Necessity is in the mind, not in objects.

(c) The simple view of objects given us no idea of power or necessary connection, i.e. no knowledge of one 'object' on its own allows causal inference to another 'object'.

(d) Causal knowledge and inference, and the idea of necessary connection, arise purely from the experience of constant conjunction.

(e) This experience neither reveals nor produces any necessity in the objects; that is, it does not provide materials for any rational inference from cause to effect (or vice versa) in a new instance.

(f) Belief consists in the liveliness given to an idea by association with a present impression.

(g) Cause and effect are distinct existences, and the ideas of a cause and its effect are distinct.

(h) A cause is conceivable without its effect and vice versa.

(i) There can be no demonstrative proof that every event has a cause.

(j) Specigic attempts to prove that every event has a cause fail.

(k) Npo impression of power or necessity is supplied by any observed relation or sequence.

(l) Causal knowledge is acquired in cases where we experience constant conjunction.

(m) The problem of induction dilemma.

(n) Belief is not a separate idea but a feeling.

(o) Belief is produced by association.

Then what is causality to Hume? In a nutshell: it is a relation between two, different, events, occupying different moments in time, and space, constantly linked. The latter is responsible for the impression that a necessary connections between them exists, while the sole fact of the observed constant connection, is sufficient reason for concluding that a causal connection between two events exists.

The major role played by the perception of the constant connection between events, points to the importance of mind in Hume's theory of causation. Forming of judgements about causality seems based mainly on our knowledge of such constant connections. Human mind forms those judgements based on the system of belief, stemming from our knowledge of the constant connections between events, and in turn, we reason based on both observed, as well as unobserved, regularities. Those play a role in building up our system of belief. The difference between causation and correlation exists only psychologically, and it makes no difference, in the world, if the relationship we identify as causal is only correlational, as causation is only appearance.

Is such an approach to causality relevant in the current discussion on the problem of free will? The most striking parallel can be drawn here between Hume and Frankfurt's argument refuting the principle of alternative possibilities, where said possibilities do not have to exist in the world. An action is free when the agent had reasons to perform it, regardless if he could have done otherwise, therefore the existence of choices does not matter. On Hume's account the question is causation exists seems to, at least from the perspective of an agent, be of a similar nature: we perceive certain correlations, and that is what counts. We can assume that those are backed by laws, but there is no way to be sure. Therefore, for the discussion on free will, Hume's approach might fit, if our interests are not in providing a proof for genuine causal links between events (ones that would exist in the world), but for perceived causal links, as in Frankfurt cases, where having reasons is sufficient for free will, regardless if those reasons were causal. Hume's approach to causality seems then to be shared by compatibilists, for who it is sufficient that we perceive our intentions, or reasons, as causal, for our actions to be considered free (e.g. Dennett).

Mackie and complex regularities. Necessary and sufficient conditions vs. INUS conditions.

Often, when discussing causality, singularity of the notion seems to be assumed: one effect has one (necessary and sufficient) cause. Although, as was briefly mentioned in the introduction to this chapter, the two concepts of necessary and sufficient were used since J.S. Mill, their exact definitions are often ridden with ambiguity. This is particularly true about the notion of sufficient cause.

Some authors (e.g. Anscombe, 1971) argue that one concept of causation is simply not enough, and different aspects of it might be relevant in different contexts. Causality might be then, rather than as Hume argues, a set of regular co-occurrences, a complex set of conditions that have to be satisfied instead. Mill himself was advocating the rejection of the assumption that one effect has only one cause and that it can be produced in only one way (Mill in *System of Logic,* Book III, Chapter 10, Section 1). Rather he argues that the same effect might be produced in different ways, can follow different antecedents etc., thus the need for distinguishing necessary and sufficient conditions. Such approach was most notably developed by Hart and Honore (1959), and Mackie (1965, 1974).

Mackie's position, based on the Hart and Honore's, emphasizes the multiplicity of causes that an event has, or might have. The cause and the effect additionally appear in a set of background conditions, which are themselves taken to be relevant. Therefore Mackie, following Mill, argues that the cause, instead of being one specific event, is rather the sum of both the positive and the negative conditions (Mackie, 1974). This can be illustrated by an example, used by Mackie, of a house fire. Here we could identify the cause as a, for instance, short circuit, but additional events also could have played a role, which include e.g. presence of flammable materials nearby, the presence of oxygen etc. Most of them, far from being both sufficient and necessary conditions for the fire, still contributed to it causally.

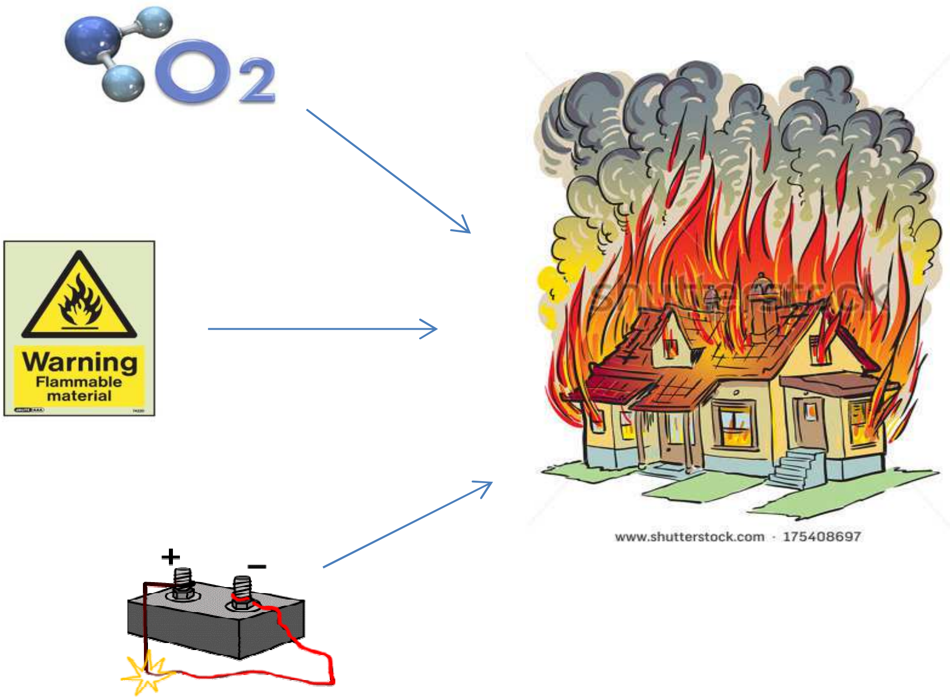


Figure 7. Why is the house on fire?[8]

In, for instance, a context of an investigation to a possible arson, an expert might identify a short circuit, as the one cause of the house burning down, while in fact the cause identified by an expert, was a part of a complex set of conditions, of which it was an indispensable part. Other conditions, like the presence of the oxygen, or the flammable materials placed nearby the short circuit, while omitted in the investigation, seem just as indispensable, in the end the house would not have burned down if the oxygen was missing.

Unlike on Hume's account, for Mackie causation is reflected through the complex, rather than the singular regularities, causal statements cannot be properly analyzed based on

[8] Source: house on fire: https://www.shutterstock.com/image-vector/house-on-fire-175408697; oxygen: http://www.generationecologie.net/2016/10/pourquoi-le-taux-d-oxygene-dans-l-atmosphere-a-diminue.html; flammable warning sign http://www.safelincs.co.uk/photoluminescent-flammable-warning-signs/; short circuit: https://c03.apogee.net/contentplayer/?coursetype=foe&utilityid=wppi&id=4701

singular statements. Those complex regularities, that potentially play a role in the bringing about of the effect, or INUS conditions, are defined as:

[…] insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result. (Mackie, 1965, p.245),

and in more formal terms:

A is an INUS condition of a result P if and only if, for some X and for some Y, (Ax or Y) is a necessary and sufficient condition of P, but A is not a sufficient condition of P and X is not a sufficient condition (Mackie, 1965, p.246)

But what does that exactly mean? Mackie illustrates the INUS conditions in action, referring to the fire example illustrated above:

*Let A stand for the INUS condition- in our example, the occurrence of a short-circuit at that place- and let B and C' (that is 'not-C', or the absence of C [*Note: the notation used by Mackie slightly differs]) *stand for the other conditions, positive and negative, which were needed along with A to form a sufficient condition of fire […]. Then the conjunction 'ABC'' represents a sufficient condition of the fire, and one that contains no redundant factors; that is 'ABC'' is a minimal sufficient condition for fire.[…]Now provided that there is some necessary and sufficient condition for this result, the disjunction of all the minimal sufficient conditions for it constitutes a necessary and sufficient condition.* (Mackie, 1965, p.246)

Therefore: the short circuit and the presence of flammable materials along with oxygen, all create a minimal sufficient condition for the fire of the house, in the specific circumstances above. The set ABC', a minimal sufficient condition for an event, is then not the only one, but rather one of several possible sets of minimal sufficient conditions. Therefore, besides the short circuit, flammable materials and oxygen, there might be other factors, that worked as a cause in starting the fire of the house. Those taken disjunctively constitute necessary and sufficient conditions; although it is possible that there are events P with no necessary conditions, but with a set of minimal sufficient conditions. In such situations they would not constitute INUS conditions. What seems immediately problematic here is the amount of information that can be extracted from such statements. They seem to be always context dependent, where the explanatory goal could heavily influence the types of causes we will take into account. Moreover we might only ever get a very limited understanding of what it is for an event, or set of events, to cause another event. To clarify this, Mackie argues that when we make a statement of a form ‘A caused P’, we implicitly have several claims in mind, namely:

(i) *A is at least an INUS condition of P […]*

(ii) *A was present on the occasion in question.*

(iii) *The factors represented by the ‘X’, if any, in the formula for the necessary and sufficient condition were present on the occasion in question.*

(iv) *Every disjunct in ‘Y’ which does not contain ‘A’ as a conjunct was absent on the occasion in question. (Mackie, 1965, p.247)*

Mackie does not claim that this is an exhaustive list of what is meant by causation, or even by the sentence’ A causes P’, but rather that this is at least a part of what is meant by any of the notions at any given time. The multiple, different causes, create what he calls the causal field. The notion itself was introduced previously by Anderson (1938), in connection to Mill’s concept of causality, and was illustrated by Mackie using influenza.

Asking about the causes of influenza can have different meanings. Usually what we think is: what causes flue in human beings in general, and in order to answer a question of this type, we can conduct a series of experiments, involving humans in various conditions, contracting, or not contracting, influence. The causal field here consists of human beings in general. The same question can have another meaning where, given the presence of the flue virus, we wonder why some people get sick, while others don’t. In this case the causal field is: the human beings in conditions where the flue virus is present. In both cases the cause is used to differentiate between conditions where, in this case, the flue occurs and those in which it does not. In the case of singular causal statements, when we ask what caused his flu, we may mean, why he, and not his wife who was exposed to the same conditions, got the flue. In such cases the causal field would consist of the people exposed to the conditions with a flue virus present. In complex cases however, such as in the house-burning example, the causal field seems to be chosen somewhat arbitrary, which Mackie himself admits is one of the shortcomings of this approach.

The causal field should not be mistaken for the set of causes, but rather should be interpreted as:

[…] a background against which the causing goes on. (Mackie, 1974, p.63)

He proposes to, given the causal fields, redefine causal statements of the form A caused P into A caused P in relation to the field F, where A is at least an INUS condition of P in the field F. This, although again to some extent arbitrary, and often not very precisely defined, allows for, at least, rejecting some of the clearly irrelevant potential causes, as they are not included in the field under consideration.

The causal field approach was criticized by van Fraassen (1980), who argued the number of the causes of an event is equal to the number of explanations that can be found for that specific event. Those explanations are simply pragmatic, and do not have to fit the real world as a whole, and here again a parallel can be found between this explanation and some compatibilist positions (e.g. Dennett), who argue that we are not in the position of Lapleacean demon, therefore the knowledge of the totality of states is not available to us. Similarly here, the explanations are pragmatic and reflect the questions relevant to the specific users, to which the total knowledge of all the background conditions is unavailable. Moreover the causal links are not chosen through the analysis of individual preferences and questions, but rather by referring to the competing counterfactuals, and what we choose to manipulate and what we choose to hold fixed. However, as we will see later on, in the discussion on interventionism, this is also problematic. Van Fraassen’s position seems clear on two points: causality depends on the context because of it counterfactual nature, and there is no one, objective, way to evaluate which one of the causal link is the real one.

On Mackie’s account more than one INUS condition is possible, thus an event could have more than one cause. Most of the time, both in the singular as well as in the general causal statements, more than one cause will be possible, also after the causal field has been determined. Those different causes can then be divided onto, for instance, predisposing and triggering causes, or main and secondary, but most of the time they will all be non-redundant. Therefore there will be a range of possibly relevant factors, and it would be implausible to expect that one of them will be the necessary and sufficient condition (if there will be such an event at all). The formula that would cover replace such condition would consist of:

[…] a disjunction of conjunctions in which each conjunct is a single term or a negation of one, that is, a formula such as "(ABC or GH or J)". (Mackie, 1965, p.255)

The possibility of multiple causes where the necessary and sufficient conditions are identified through a disjunction of conjunctions, as we will see further on, could be of particular relevance in connection to the free will debate.

One possible problem that can appear in connection to the multiple causes of one effect might be the different weights, or importance of different causes. Although the causes are different, some of them might also be deemed less important than others. As in the example of the burned down house, the presence of the short circuit can be considered more important than the presence of flammable materials nearby, and ever more so than the presence of the oxygen in the air. In the end the oxygen is always present, and the houses do not burn down because of that. Mackie however argues that the various INUS conditions are equal, therefore:

If two factors are each necessary in the circumstances, they are equally necessary; one inus condition cannot be more of an inus condition than another. (Mackie, 1974, p.128)

This could also be due to the specificity of each circumstance: while oxygen is always present, in the specific case when it is combined with the short circuit, it becomes a factor contributing to a fire. However, although the different INUS conditions are equally necessary, some of them might be easier to replace than others, and exchanging some might have an influence on the effect, while others less so, which remains a problem.

As for free will, intuitively speaking, it seems that for a free action might be one caused by a number of different causes. As we will see later on (in the chapters on what is causal in action generation, and on mental causation) some of the problems connected to free will, understood in terms of being caused 'in the right way', could be seen as originating from thinking that there is one, necessary and sufficient, cause for a specific action.

Anscombe and Davidson on causation.

Another way to remedy the issue of necessary and sufficient condition is to argue that, instead of a one general concept, we are dealing with a number of different causal concepts, which are described as causes. That general description is misleading, as the more detailed descriptions of the relationship are much more informative, and, moreover, are not synonymous with one another. This is the approach of Anscombe, whose work on causality is closely related to her views on intentions and reasons, all relevant to the free will problem.

In this section I will discuss Anscombe's work along with the arguments presented by Davidson, and, while their positions are not similar, the reason for it is their involvement in the causalism/anti- causalism debate. Both positions are taken to be classic, and they will be discussed in the forthcoming sections of this dissertation, therefore is makes sense to also present their views on the notion of causality together.

Anscombe.

Anscombe's most notable contribution to the debate on causality comes from her article, and inaugural lecture, *Causality and Determination* (1971). In her work, she argues against strictly associating causality with necessity, and seems to think that, often, the knowledge of the cause, rather than being predictive, is derivative. Thus causes are derived from their effects, rather than investigated and discovered first, and used to predict the occurrence of the effect. Also the effects could be derived from their causes. This could potentially suggest similar intentions as in Mackie's approach, since also there what we identify as a cause is contingent on the effect (e.g. if we investigate why the house burned down, the short circuit will be the cause most relevant to us). Along with the rejection of the necessitation, Anscombe also argues against the humean approach of the constant conjunction between the cause and effect.

In Anscombe's work a major role is played by the notion of derivatness, which can cover both production of the effect by the cause, as well as making a difference to the effect. Therefore Anscombe's work is compatible with both the production as well as difference making approaches. She uses an example of physical parenthood to illustrate her point where she states:

[…] everyone will grant that physical parenthood is a causal relation. Here the derivation is material, by fission. Now analysis in terms of necessity or universality does not tell us of this derivedness of the effect; rather it forgets about that. For the necessity will be that of laws of nature; through it we shall be able to derive knowledge of the effect from knowledge of the cause, or vice versa, but that does not show us the cause as source of the effect. Causation, then, is not to be identified with necessitation. (Anscombe, 1971, p.92)

While the position is compatible with both views on causality, mentioned above, the notion of derivation of causes, can suggest that production would be closer to her intuitions. Seeing the cause as the source of effect can be red as meaning that the latter is produced, in some sense, by the former. However, the how of it, is not covered by seeing causal relationships in terms of necessity connected to the laws of nature. The connection between necessity and causation can be then described as follows:

[…] a cause C is a necessitating cause of an effect C when (I mean: on the occasion when) if C occurs it is certain to cause E unless something prevents it. C and E are to be understood as general expressions, not singular terms. (Anscombe, 1971, p.101)

Here again we can see that causation is only at a general level, while singular terms, describing events connected in a more specific manner, are more detailed and more informative. Therefore we can conclude that a duality exists: from one angle causality only exists on general level, as an umbrella term for the singular, more specific terms describing the connection between the cause and the effect. The existence of causal links is then not covered by laws of nature, nor coming from the observation of constant connections, and moreover, identifying that the two events are connected through causality, does not mean that this connection is necessary.

Anscombe's rejection of the laws of nature as a basis for identifying causal links, and governing causality, is discussed in connection to indeterminism, where, in the deterministic universe, the predictive power of causation is evident. This disappears when we bring indeterminism to the table. The two natures of the universe are compatible with two different natures of causality, and necessitation entails determinism. Since the two are only theories, and the derivative nature of causality is compatible with both, it might just make a better theory to assume the latter. On determinism Anscombe writes:

I am not saying that a deterministic system involves individual predictability: it evidently does not. But in default of predictability the determinedness declared by the deterministic system has got to be believed because the system itself is believed. (Anscombe, 1971, p.96)

The question of determinism and predictability gains on significance when seen in connection to the free will debate. As we have seen in the previous chapters, the seeming consequence of determinism, predictability, is a major problem for the possibility of free will. Like Dennett, also Anscombe questions the true meaning of the word 'determined', and she postulates a certain disconnect between the two notions: causality and determinism. She emphasizes that a thing hasn't been caused before it had actually occurred, while it might have been determined before it actual came to be. Describing an effect as being determined, means for Anscombe, that from the range of possibilities of the antecedent, we allow only one of them to have the potential to become true. If we see this through the free will discussion, it could serve as an argument against the original formulation of the principle of alternate possibilities, as it leaves little room for those different action options to exist, however is still compatible with the revised, Frankfurt's version of it. For Anscombe, the meaning of the word 'determined' is not necessarily linked with orienting ourselves to the past, looking and past events, which usually leads to having no alternatives, because the events have already taken place. Being determined then means we are able

[…] to give content to the idea of something's being determined, we have to have a set of possibilities, which something narrows down to one-before the event. (Anscombe, 1971, p.98)

This approach to determinism, as oriented to the past, combined with Anscombe's rejection of the causality and predictability of it (a thing has not been caused before it happened, and it is only determined when it already happened), could have the potential to solve some of the issues identifiable in the classic debates on free will. The different possibilities could be seen as disconnected from the issue of causality, and it could allow indeterminism into the debates, since even for an event to be determined, there needs to exist a set of possibilities, getting narrowed down to one, determined one.

The derivative nature of causal claims is significant also in connection to the neuroscientific discussions on the conscious decision. It leaves the possibility that the moment of choice, perceived by the participant, is also just derived from the action itself. Since, e.g. on Libet's set-up, we have to rely on participant's reports, it is possible that the moment of conscious decision, seen as a cause, is just derived from the action. This does not necessarily mean bad news for free will, as for the physical underpinnings of our action, Anscombe finds it implausible that the same train of physical events would always result in the same action. Therefore the decision would still be seen as having some causal impact, since causation, as production of the effect, and causation, as explaining what happened, seem to be very different on Anscombe's account. The former, indeterministic, while the latter, linked closely to determinism.

Davidson on causation.

Unlike Anscombe, Davidson holds, after Hume, that the correct theory of causation involves some form of the regularity approach, and the events taking part in the causal relationship are individuated according to their position on the time-space continuum (Paul, 2004, p.209), again emphasizing the temporal order of the cause and effect. Following Hume he not only argues that causation is somehow connected to the co-occurrence of the two events, but also that some laws have to govern the relationship between causes and effects. Those causal laws:

[…] differ from true but nonlawlike generalizations in that their instances confirm them; induction is, therefore, certainly a good way to learn the truth of the law. (Davidson, 1980, p.18)

Therefore, the observations of constant connections are sufficient to determine that two events are in fact parts of the causal chain, governed by some laws. Davidson does not claim that this is the only way to learn the truth of the laws, and additionally, in order to know if the singular statements covered by laws are true, we do not have to know if the law itself is true. It is sufficient to know that such a law exists.

Davdison, unlike Anscombe and Mackie, believes that the necessary and sufficient cause can be identified. He argues that sentences such as *the cause of this match's lighting is that it was struck* (Davidson, 1967, p.81) constitute a description of a cause, and, as we add to it, we can reach a point where we will be able to deduce from the description, and some laws that govern this specific connection, if the relationship is causal. For Davidson:

The fuller we make the description of the cause, the better our chances of demonstrating that it was sufficient (as described) to produce the effect, and the worse our chances of demonstrating that it was necessary; the fuller we make the description of the effect, the better our chances of demonstrating that the cause (as described) was necessary, and the worse our chances of demonstrating it was sufficient. (Davidson, 1967, p.82).

Therefore a trade-off between the description of the cause and the necessity of it, and between the description of the effect and the sufficiency of the cause are identified. Linking to Anscombe's position, the various causal links can have more or less specific descriptions that influence the way cause, and effect, are identified. However, for Anscombe, the descriptions can only be specified on the individual level of events, which does not seem necessary for Davidson. Nonetheless, the proper identification of the causal link seems to, in both cases, depend on how exact and full the description is, but on Davidson's account, if the description is vague, we can rely on the laws that govern such relationships. The only, or the main, reason for accepting singular causal statements is the dependence of their truth-value on the existence of the covering laws. It is also important to remember, that for Davidson the events can be differentiated only based on the way they are described. As we will see further, in the chapter on mental causation, the difference between the mental and the physical, lies in the type of vocabulary used. The mental is described in mentalistic terms, the physical is described in physical terms, nothing else, like a different mode of existence, is necessary, which shows the importance language plays in Davidson's approach. On the level of tokens all events are physical, only on the type level, where descriptions are essential, the events can be distinguished from one another.

All three, Davidson, Anscombe and Mackie bring to the attention, in different ways, a crucial problem in the identification of causal links, namely that one effect could have multiple causes, and that identification of the one, necessary and sufficient, might be more problematic than initially assumed. In the scenario where an event has two, or more, causes, depending on the assumed possibility of finding one necessary and sufficient cause, those can be seen as either competing, or complementary. This division, into necessary and sufficient one cause, on the one hand, and multiple causes on the other, suggests that any discussion on what causation is, and how things are caused, has to first clarify which answer they are after: are we looking for one cause, or rather a causal field. Are the co-occurences sufficient to determine the existence of a causal link, or something more potent is necessary. All of these answers can have major consequences to the free will debate, as discussed in connection to the causes of our actions.

Counterfactuals.

Counterfactual theories of causation hold that the causal relationship can be explained by referring to counterfactual conditions of the form *If X had not occurred, C would not have occurred,* emphasizing the counterfactual nature of causal links. Here two different frameworks can be identified: (1) we are committing ourselves to the view that causation is that counterfactual dependence, or (2) we are suing counterfactual analysis to determine the existence of causal links, in terms of assessing possibilities, while defining causality itself as something other than counterfactual dependence. The latter approach has been used throughout the modern discussions on the topic, most notably by interventionists, who, while holding that causality is difference making, often employ counterfactual analysis. The same approach can also be found in neuroscientific experiments, although here it is not clear what the view on causality clearly is.

Counterfactual analysis of causation was already introduced by Mackie, who argued a cause was something that made a difference in a specific context, emphasizing also the contextual nature of causality. Still, the most famous counterfactual account to causality can be attributed to D. Lewis (mostly in his 1973(a) and article, 1973 (b)). Here, conscious of space and time constrains, I will present only a limited account of the counterfactual position on causality, and, whenever I will refer to a position as being counterfactual, I will mean that it conducts the analysis of causation using the counterfactual conditional of the form quoted above; if the meaning of the word used is broader I will emphasize that. Therefore I will treat approaches as counterfactual if they, in their analysis of causal claims, consider counterfactuals, hypothesizing what would have been the case if the world would have been different. Example of solutions of this form was already given throughout this chapter and the previous ones, where I discussed Lewis's solution to the alternate possibilities problem. Importantly, some of the positions discussed here are not easy to clearly categorize as either regularity or counterfactual theory, even in Hume traces of the counterfactual analysis can be found, the same can be said about, already mentioned above, Mackie, whose account has been classified as both a regularity theory, as well as in line with Lewis approach. Although the exact categorization is not of major importance, the difficulty is worth noting.

In Lewis' theory, the counterfactual dependence that holds between two evens, the cause and effect, is enough to classify such link as causal. Therefore, in a nutshell, we can conclude that an event E is a cause of an event C when, in the close possible worlds, if the E does not appear, also the C does not.

Similarly to Mackie, Lewis does not argue an effect has only one single cause; on the contrary, he recognizes that sometimes we choose an event as <u>the</u> (necessarily and sufficient) cause based on rather mysterious criteria, while ignoring other causally relevant events. However he does not explore those criteria further, arguing that those do not lie in the scope of his interests, focusing rather on the question what it is to be one of the causes (Lewis, 1973 (b)). Performing a counterfactual analysis is based on an investigation of the various possibilities, through modelling the possible worlds, where:

[…]the prevailing laws of nature are such that there do not exist any two possible world which are exactly alike up to some time, which differ thereafter, and in which those laws are never violated. (Lewis, 1973 (b), p.196)

In the quote here, Lewis is defining determinism, through emphasizing the importance of laws of nature is visible. The crucial step of Lewis's counterfactual analysis is the notion of possible worlds and their proximity to the actual world. Those similarity conditions between the possible worlds set the truth value of the causal statement, or its truth conditions in other words. The closer the possible world to the actual world, in colloquial terms, the better, and the world is seen as closer to actuality, where

[…] if the first one [the possible world] *resembles out actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off against one another.* (Lewis, 1973(b), p.196)

The similarity between the worlds is judged on the basis of all aspects, both the laws of nature, as well as the facts. As we have seen previously, Lewis considers laws of nature as paramount (e.g. breaking the laws in order to guarantee alternate possibilities), and breaking them should be avoided. However, in some cases, where the worlds are very similar in other aspects, but the laws of nature between them differ, those worlds could potentially be still considered close. Comparisons are done both with regards to but the actual world, as well as across possible worlds, called by Lewis comparative similarity.

The counterfactual dependence can be described as the set of truth conditions for the counterfactual: if A was the case (or was true), C would have been the case (or would have been true), can be defined as being true, when either:

(i) *There are no possible A-worlds, or*

(ii) *Some A-world where C holds is closer (to w) than is any A-world where C does not hold.* (Lewis, 1973b., p.197)

Therefore for the counterfactual to be (non-vacuously) true, the world, or worlds, where the consequent is true along with the antecedent are closer to the actual world, than the worlds where the antecedent is true without the consequent.

The causal dependence, is constituted when the occurrence of some event A is dependent on the occurrence of another event B, considered the cause. Such causal dependence implies causation. As Lewis puts it:

If c and e are two actual events such that e would have not occurred without c, then c is the cause of e. (Lewis, 1973b., 200)

Causation however, unlike causal dependence, is transitive, therefore the converse of the statement does not hold.

Lewis's approach to causation has become the standard, with his analysis being used also in accounts that do not commit themselves to equating causality with counterfactual dependence. The usage of the possible world analysis can also be seen as useful in the interpretation of the neuroscientific experiments, based on manipulation, where the connection between the cause and effect is defined by such dependences

The law- based understanding of causality, currently, is not the dominant one anymore, rather approaches focusing on the specific relationship between the cause and effect, are often dominant. In the next chapter I will analyze some of those, again, choosing the positions that can be seen as immediately relevant for the larger problem I want to discuss here (mental causation). Because the two chapters (chapter four and five) have the same focus, just cover different positions, I have chosen to, unlike in other chapters, only write one conclusions for both. Therefore the conclusions of chapter four, are given at the end of chapter five, together.

Chapter 5. Modern approaches to causality.

The previous chapter has been concerned mainly with explaining the most basic concepts connected to the issue of identification of causal links between events. I have discussed the various ways to approach the topic, from questions regarding causality itself, to investigations into causal explanations. I have also discussed the more classical approaches, like Hume, Davidson, Anscombe and Lewis, and their modifications, like Mackie's INUS conditions.

I have shown how, on Hume's account, all there is to causation is causal regularity, as the only way we can acquire any knowledge about causality is through observation of those regularities, which in turn can only be granted limited trust. Without experience we would not be able to see any causal links, or causal regularities, but induction is not sufficient to allow us to conclude that those regularities will always happen, in the future. Causality is covered by laws of nature, yet those are inaccessible to us.

Mackie's theory follows Hume's in its core assumptions, but questions Mill's notion of necessary and sufficient cause, arguing that for any given effect, a number of causes can be identified. Moreover, in the process of identification of those causes, both the context in which the evaluation of what is the cause, as well as the context of the occurrence of the effect matter. This was explained on the basis of the burning house example, where both the context of the fire, as well as the context in which the cause was determined (e.g. fire insurance investigation), impacted what was identified as a cause. To avoid this relativity, instead of speaking of one, necessary and sufficient cause, Mackie focuses rather on the causal field, where INUS conditions are investigated.

Anscombe and Davidson, both emphasize how important the descriptions are, in the discovery of the causal links between events. This more of less exhaust the similarities between them, as Davidson follows Hume's approach, while Anscombe presents an opposite view, where causation is seen as derivative from the effect. Moreover Anscombe emphasizes the generality of the notion of causality, preferring more domain specific descriptions, that, according to her, are significantly more informative.

Lewis counterfactual approach departs from Hume's position, and sees causality as counterfactual dependence, introducing the possible world analysis, which has become a standard in the current discussions on causality.

In this chapter I will focus on the modern theories of causality. Also here the discussions are concerned not only in defining causality itself, but also finding ways in which the causal links can be identified, and described. I will discuss the approaches immediately connected to the free will debate, either through the problems they present to it, or the solutions they propose: the production accounts and the difference- making accounts. For the former causality is understood as some form of productive process linking two counterparts, while the latter emphasize that being a cause is making a difference to the effect. Both the production, as well as the difference making, accounts make use of the counterfactual analysis mentioned in the previous chapter, so in some aspects their understanding of the dynamics of the causal analysis is similar (although e.g. Woodward, 2003, p.132 argues that Salmon's account in particular aims to distance itself from the standard counterfactual analysis).

Processes.

The parts of the causal links, in many causal theories of action, are taken to be events. As we have seen previously, for the free will, this might cause problems, such as the missing agent issue, discussed in previous chapters. Moreover, in connection to neuroscience, although currently the dominant approach is also to take various events as part of the causal chain, another approach could be more promising. This could involve, rather than just singular events, processes with interacting parts.

Such understanding of the causality, as process, where different parts of the larger whole, which is the process itself, was initially developed by Salmon who thought that causal processes as well as causal interactions can be seen as more fundamental than the causal relationships that can be seen occurring between events (Dowe, 2009, 2012).

The processes themselves are understood here as a series of events connected together through some characteristic, we call causation. The recognition of connection, or either world lines or causal lines, can be traced to Russel (1948), who was first to describe a causal line as series of events, connected in such a way that, given one of those events we can conclude some facts about other events, without knowing anything about the environment in which they exist. Russel's theory however did not allow for a clear distinction to be drawn between actually causal processes and pseudo processes, which created an obvious problem, which Salmon's account was supposed to be a solution to. Salmons' work is the first in a series of different solutions with processes being central, most notably found in the work he conducted with Dowe, and these two accounts will be the ones I will focus here.

Salmon and Mark Transmission.

The first of the process accounts, and rather quickly dismissed is Salmon's mark transmission. Since the problems of the account have been well covered in the literature, here I will only present a sketch of the position, quickly moving to the account developed later by Salmon and Dowe. Most detailed account of mark transmission account was presented by Salmon in his book *Scientific Explanation and the Causal Structure of the World* (1984), where he describes a causal structure of the world as consisting of a network of processes, between which, in case they are causal, a mark is transmitted.

Let P be a process that, in the absence of interactions with other processes, would remain uniform with respect too characteristic Q, which it would manifest consistently over an interval that includes both of the spacetime points A and B [where A is not equal to B]. *Then a mark (consisting of a modification of Q into Q^*), which has been introduced into process P by means of single local interaction at a point A, is transmitted to point B if [and only if] P manifests the modification Q^* at B and at all stages of the process between A and B without additional interactions.* (Salmon, 1984, p.148)

Here the initial process is proceeding without change, until a modification in one of its characteristics is introduced. If that modification gets transmitted between two points belonging to the process, without any other modifications being introduced, or taking place, the former can be called causal. Dowe (1992) describes Salmon's general objective as a construction of a theory in which causality would be an objective and contingent feature of the world, while being compatible with both determinism as well as indeterminism.

Mark transmission however, as was quickly pointed out, failed to distinguish between causal and pseudo causal processes. As a consequence, a shadow, on mark transmission theory would be considered a causal process. To illustrate this: a moving car can be considered a process, just like a shadow of a car. Applying the definition above, given that the shadow accompanies the car from the beginning till the end of its journey (assuming that the car is actually going somewhere), it should be considered causal. Obviously this conclusion would be mistaken, and a shadow is not a causal process, however since there are no clear conditions on the types of marks, which could be considered as causal, there is no way to discredit it from being causal.

Those issues, recognized by Salmon, led to largely abandoning the approach, and developing, with Dowe a new type of process account, centered around conserved quantities, where the shortcomings of the previous account should be avoided.

Conserved Quantity Theory.

The first sketch of the conserved quantity theory was presented by Dowe (1992), as a result of an analysis of the problems identified in connection to the mark transmission approach, as developed by Salmon. Dowe presents two main definitions of causal interaction and a causal process where:

DEFINITION 1. A causal interaction is an intersection of world lines which involves exchange of a conserved quantity.

DEFINITION 2. A causal process is a world line of an object which manifests a conserved quantity. (Dowe, 1992, p.210).

Unlike in the case of Salmon's approach, where the mark was a very vague notion, the conserved quantity is a term well-grounded in physics, and therefore less ambiguous. Therefore the notion is taken to mean *any quantity universally conserved according to current scientific theories (*Dowe, 1992, p.210), which includes mass, energy, momentum and charge. By the world line, mentioned in the first definition, Dowe understands a collection of events on a space-time continuum, which represents the history of an object. An exchange of the conserved quantity happens on the intersection of at least two, but possibly multiple, meet. From the two at least one has to be an incoming, and one outgoing process, and the connection is conserved causal when a conserved quantity has been exchanged between them, therefore both processes share some value of, for instance, energy or momentum.

The Salmon and Dowe's approach allows for a very scientific account of causation, closely linked to the scientific practice. The identification of the causal connections does not rely on inaccessible laws of nature, but rather is based on quantities defined in physics. In a sense their proposal makes the problem of causation into physical rather than metaphysical problem. However, while being grounded scientifically improves the disambiguation of the terms used, at the same time it limits the applicability of the account. Most notably, while working well when applied to the physical system and operated within the physical theories, it seems to fail when applied to biology and to social sciences (the work of Marchamer, and Russo respectively). Moreover it fails to account for cases in which causal links are usually considered present, yet no conserved quantity is exchanged. Paradigmatic of such examples can be illustrated by a sentence 'I was late to work because the bus did not arrive'. The so-called problem of absences describes situations in which there clearly is a causal connection, but the cause is absence rather than the presence of an event or a cause in the usual sense of the notion.

Why are these approaches relevant to the discussions on free will? Firstly, as we will see in the proceeding chapters, if we aim to argue that, in order to be defined as such, a voluntary, intentional and free action has to be caused in 'the right way', and that way is via a mental event (as I will show). If that is the case one of the biggest problem we will have to deal with is the Exclusion argument, constructed specifically with a productive understanding of causality in mind. Moreover it seems that neuroscientific approach favors accounts which permit explanations in terms of physical, transmitted quantities. This is understandable, as the brain operates through propagation of the chemical and electrical signals. However, this could lead to further issues, as we will see in the chapter on mental causation. Additionally, the great advantage of these accounts is their compatibility with both determinism and indeterminism. As we have seen time and time again here, any account that does not exclude or favors one over the other, would be preferred, as I am aiming to stay agnostic on the issue. Additionally, since I want to argue that causation, and especially mental causation are a core condition of existence for free will, I have to argue that causality is something that exists in the world, rather than is just a mode of description. Therefore a theory in lines with those general assumptions would also, I hypothesize, be beneficial for the solution of the free will problem.

Mechanisms.

Another production type accounts, this time involving mechanisms rather than processes, has been proposed by, among others, Craver (2007). His position is particularly interesting when applied to the question of the causal origin of action, as it is linked with causality as used in neuroscience and biology. Also the interest in mechanisms in philosophy of biology and neuroscience has been around for some time now (Bechtel and Richardson, 1993; Craver and Darden, 2001; Marchamer, Darden and Craver, 2000).

Similarly to the process accounts, also in the mechanistic approach what matters most seems to be causal explanation, rather than the nature of causation itself, and the former is then used to account for the latter (in a nutshell: causality is a mechanism). The account proposed by Craver in his 2007 book, and in his work with Bechtel, is the most exhaustive account of the mechanistic approach to causality, applied specifically to neuroscience, therefore the most interesting in the context explored here. Of course the notion extends far beyond biology and neuroscience, including the most classic examples of machinery, however, as Craver argues equating mechanistic causality with a notion of a machine is too restrictive to account for the advanced neuroscientific processes. Therefore, as a paradigmatic example of what he means by mechanistic causality Craver uses the example of a neuron releasing the neurotransmitter (Craver, 2007, p.4), where:

The mechanism begins, we can say, when an action potential depolarizes the axon terminal and so opens voltage-sensitive calcium (Ca^{2+}) channels in the neuronal membrane. Intracellular Ca^{2+} concentration rise, causing more Ca^{2+} to bind to Ca^{2+}/Calmodulin dependent kinase. The latter phyospohorylates synapsin which frees the transmitter-containing vesicle from the cytoskeleton. At this point Rab_3A and Rab_3C target the freed vesicle to relase sites in the membrane. Then v-SNARES(such as VAMP), which are incorporated in the vesicle membrane, bind to t-SNARES (such as syntaxin and SNAP-25), which are incorporates into the axon terminal membrane, thereby bringing the vesicle and the membrane next to one another. Finally, local influx of Ca^{2+} at the active zone in the terminal leads this SNARE complex, either acting alone or in concert with other proteins, to open a fusion pore that spans the membrane to the synaptic cleft. (Craver, 2007, pp.4-5)

Mechanism understood in such way if centered on both the parts of the mechanisms, the entities, as well as all the activities of those parts, which are organized and explain the phenomenon in question. Therefore the mechanistic approach will cover both the entities (e.g. neurons), and their activities (e.g. firing) in order to explain the complete phenomenon (e.g. transmitter release). This approach can then potentially be extended to account not only for the very basic entities, like neurons, but potentially the totality of agent's behavior. Entities are understood as components of mechanisms, which have the ability to engage in a number of versatile activities, are spatiotemporal, with a specific size and a structure, have masses, charges and momentum (Craver, 2007, pp.5-6).

Activities, on the other hand, are the 'causal glue' in the mechanisms, and are considered to be productive therefore they are not only correlations, and they can be exploited for the purposes of manipulation. This does not entail that mechanisms are in line with difference-making accounts of causality, rather they use manipulation, and counterfactual dependence, to show how causality propagates through the mechanism. The underlying intuitions, in the mechanistic approach, seem to be that the effect is somehow produced by the cause, and the causal link itself is something more than mere difference making. Activities together with entities organized in such way as to explain the phenomenon in question constitute causality.

The main difference between the mechanistic approach and the previous approaches to causality, such as Hume's approach, is not relying on the existence of the covering law. The link in those is shown to be causal, if the expectations regarding the result are met, as based on the antecedent conditions combined with the laws of nature. Since strict, or not ceteris paribus, laws are difficult to find in special sciences, it became questionable if those accounts can really be applied here. Both mechanistic approaches described here, as well as the previously discussed positions of Salmon and Dowe, all abandon this path and show that determining if a link is causal, requires showing how the effect is produced by the causes, therefore production accounts.

It is clear that the mechanistic account was designed with neuroscience in mind, especially since the transmission of the signal between neurons was taken as an example of such connection. The general aim of neuroscience is, or at least seems to be, to provide explanation to the behavior of organisms, through the description of the computations performed by various regions of the brain. This is certainly true about the neuroscientific approach to free will, as we have seen in the previous chapters. It also falls in line with the approach to studying free will, focusing not on the external conditions for its existence, but rather on the internal one: how it has to be caused. Therefore we could conclude that the deepest assumption here is that the action has to be produced by something, either the conscious decision, or the readiness potential (at least in Libet's approach). Additionally mechanistic approach allows for multilevel explanations, which can refer to:

[…] the behaviors of organisms, the processing functions of brain systems, the representational and computational properties of brain regions, the electrophysiological properties of nerve cells, and the structures and conformation changes of molecules. (Craver, 2007, p.9)

While important in neuroscience, the multilevel explanation is problematic, and sensitive to the causal exclusion argument, which is known to drain causal powers from all higher levels. Craver however emphasizes, that explanations in neuroscience have the tendency to oscillate between levels, and only through this we are able to explain the full mechanism of the phenomenon.

In the beginning of this chapter I have briefly discussed the difficulties one can encounter while trying to establish what neuroscientists mean exactly, when they talk about causality and causes. Because of the lack of clear definitions, we have to rely on the way the science practically explains the phenomena in question. With this in mind, mechanisms might be a good way to reflect the intuitions already present implicitly. The goal of neuroscience was described by Edelman as construction of:

[…]a scientific theory of the mind based directly on the structure and workings of the brain. By "scientific" in this context, I mean a description based on the neuronal and phenotypic organization of an individual and formulated solely in terms of physical and chemical mechanisms giving rise to that organization. (Edelman, 1989, p.8-9)

and as such seems to be perfectly in line with the general line of thinking presented in mechanistic theories.

The additional constraints, identified by Craver, that should be applied to accounts of causal connections include: (1) temporal sequences are not explanatory (if they are just temporal sequences and nothing else), (2) the effect is explained by the cause (and not the other way around), (3) two, or more, independent effects that share common causes do not explain one another, (4) phenomena that are not relevant are also not explanatory, and (5) effects do not have to be made probable for the causes to explain them (Craver, 2007, p.63).

Those constraints are also used to justify why, in neuroscience, the conserved quantity theory (explained above) is insufficient, as some of those constraints are problematic under the latter. More specifically, Craver gives two reasons why conserved quantity theories fail when applied to the neuroscientific domain: (1) [they] *do not provide a satisfactory account of causal relevance* (Craver, 2007, p.78); and (2) [they] *do not accommodate negative causal relations, such as cases of omission and prevention* (Craver, 2007, p.78).

The former problem has to do with situations where one has to establish which features of the cause are relevant to the occurrence of the effect. On the conserved quantities account this is simply impossible for some of the explanations required in neuroscience, as, besides searching for the conserved entities, one should also describe which features make a difference.

Similar problem can be observed in dealing with situations where multiple conserved quantities are exchanged, and again, on the conserved quantities account, there is no way to tell which one of them are causally relevant.

The second problem, negative causal relations, was already referred to, in connection to Salmon and Dowe's account. It concerns situations in which the cause is the lack of an event, rather than its occurrence (e.g. the bus that did not arrive). In application to neuroscience, the problem arises in explanations when, for instance, a lack of neurotransmitter, rather than its presence is a cause of a membrane activation.

While the mechanistic account seems attractive, especially since it seamlessly fits into the neuroscientific approach, it relies on the possibility of multi-level explanations, which, as I will show further on, can be problematic. This issue will be especially pressing when applied to the mind-body problem.

Interventionism or difference- making accounts.

The newest approach in the study of causality, as processes and mechanisms, breaks with the tradition of law-backed co-occurences. However, unlike in the previous cases, instead of thinking of causes as producing effects, focuses on them as making a difference to the effect. Interventionism, or difference-making, often presents this way of thinking of causality to be supported not only by theoretical considerations, but also empirically. In fact it has been broadly applied to various domains, including health and epidemiology, and the concept of manipulability is often used in sciences. Its core idea was initially described in relation to molecular biology by Weinberg (1985), who argued that the science went from being descriptive to explanatory, with the power to identify causal mechanisms, however here the sole existence of the mechanism is not leading to treating causality as synonymous with mechanisms, as on the previous account.

Interventionism exploits counterfactual, statistical dependencies, in order to investigate the nature of a relationship between two variables. The account exploits some of the basic intuitions about causality, where a cause can be identified through the manipulation of different variables. Such manipulation, when applied to the cause, should result in a change to the effect, while when applied to non-causes, should have no influence on the outcome. The position, while based on the computational models of causal inference, from the work of e.g. Judea Pearl (2000), has been most exhaustively described by Woodward, in his 2003 book *Making things happen.* The position focuses on the connections between causation and manipulation, and extends the previous work, which investigated the impact of such manipulations on the causal link. Here the manipulation itself is not only taken to be the method in which the causal links are discovered, but difference making is causality.

While closely linked to the idea that causal links can be discovered through manipulation, and treating difference making of the cause as a mark of causation, interventionism is not limited by performable experimentations, or the usage of experimental methods. Most of science is based on the idea that manipulating different variables will lead to results informative about the dependencies between them (in fact using the counterfactual conditions I have discussed previously). Those methods are applicable both in the difference making as well as in the production accounts. There, in order to identify whether mass or energy are transmitted between two entities, an experiment where we manipulate one of them, to see if leads to changes in the other, will yield the desired answers. Interventionism however, goes a step further, and limits causation to difference making. In other words, there is nothing more to a cause, in its being a cause, than making a difference to the occurrence of the effect. No 'causal oemph'.

Main definitions.

As mentioned above, while different, early formulations can be found (e.g. Gasking, 1955; vonWright, 1971), but was probably most exhaustively described by Woodward (2003). There he defines the core notions in the difference-making approach: the direct and the contributing cause as well as the notion of intervention. On his account the direct cause is defined as (Woodward 2003, p.59):

A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V.

Contributing cause on the other hand is defined as:

A necessary and sufficient condition for X to be a (type-level) contributing cause of Y with respect to variable set V is that (i)there be a directed path from X to Y such that each link in this path is a direct causal relationship (…) and that (ii) there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value.

Intervention variable (Woodward 2003, p.98) has to meet several conditions:

I1. I causes X.

I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are sucha that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.

I3. Any direct path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are build into the I-X-Y connection itself, that is, except for (a) any causes of Y that are effects of X (i.e. variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.

I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

Therefore, for the event to be identified as a cause, two conditions have to be fulfilled: the presumed cause has to be manipulable, and the other causes in the set have to be fixed. The latter condition is particularly important in the case of off-path variables, and it should prevent situations, where, while the manipulation is performed, other variable also changes, and the latter is the cause of changes in the effect. This would lead to false attributions of causality.

Another definition of interventionism that is frequently referred to, was presented by Menzies and Price (1993), although there the authors prefer to call it agent causation. In their view:

[…]the common idea to agency accounts of causation is that an event A is a cause of a distinct event B just in case bringing about the occurrence of A would be an effective means by which a free agent could bring about the occurrence of B. (Menzies, Price, 1993, p.189)

The definition offers far fewer details, than the one given by Woodward, and, moreover is much more ambiguous. Notions such as effective means, or a free agent are very unclear, and the latter notion is particularly problematic and controversial in the light of the discussions presented in this dissertation. The authors themselves, rather than about interventions, talk about means- end relationship and agent probabilities which:

[…]are thought of as conditional probabilities, assessed from an agent's perspective under the supposition that the antecedent condition is realized ab initio, as a free act of the agent concerned. Thus the agent probability that one should ascribe to B conditional on A (which we symbolize as $P_A(B)$), is the probability that B would hold were one to choose to realize A. (Menzies, Price, 1993, p.190)

Again, notions such as agent's perspective are extremely problematic in the light of the discussions on free will, and brings into foreground issues such as anthropocentricity of the account, which Woodward's position avoids, by referring to theoretical interventions. Therefore here, when referring to interventionism, I will be using Woodward's definitions, which, while still potentially assuming some form of free will, is more carefully defined.

The role of experiments in interventionism.

Crucially for the proper understanding of the position, interventions and experiments are not synonymous, but related. Experimentation, although the main way to obtain the truth value of the claims about what happens to the effect under appropriate manipulation, does not exclude that reliable inferences might come from non-experimental evidence. However, even such data, although not stemming from actual experiments, still take a form of experiments, being connected to a specific way of thinking about causality that is inherent to interventionism.

Settings, where experiments are not performed, are scenarios where the experiment could nonetheless be performed, without actually having to perform it. Some data, such as ones regarding children's performance depending on the type of school they attend (the example discussed also by Woodward, 2003, p.35) cannot be checked experimentally, for variety of reasons. First of all it would be almost impossible, and ethically questionable, to randomly assign children to private and public schools, and then monitor their academic performance. Yet it is still worth to imagine an experiment in which a manipulation of such sort would be an option, and to predict its result. Those hypothetical experiments allow to theoretically asses different initial conditions as well as a variety of possible outcomes of the interventions.

Therefore, while the two notions are undoubtedly related, the experiments do not have to be interventions in the standard sense, and can remain in the purely theoretical terms. Empirical data is not always possible to obtain, and the theoretical approach allows to conclude the truth value of the presumed causal connection (if it holds or not).

If the data from either the theoretical or the empirical experiments do not clearly support the causal connection claim, manipulation is either not well defined, or what happens under the manipulation is unspecified, the causal claim itself is unclear. In those cases the causal claims can be disambiguated by, as Woodward states, distinguishing between the variety of causal claims that can be made about the outcomes of those experiments, or by replacing claims connected to the under-defined notions of manipulation. This way of clarifying ambiguous causal statements is for Woodward an additional reason for accepting the difference making account, where the notion can be made clearer through changing what we are manipulating, or distinguishing between different, in fact independent, claims, that occur in an ambiguous causal statement.

Is causation "out there"?

As with the previous accounts, also in difference making one of the central issues is its status in the world: is there something 'out there' that we can call causality, or is it just a comfortable way to describe reality, but there is no one thing that is causality. The process accounts seem to suggest that causality exists somehow, and it is more than just a way to describe reality. For other accounts, like Hume, it seems that even though it probably does exist, it is out of our perceptual reach. For difference making, the answer to this question is not straightforward, as it depends on the author.

For Menzies and Price (1993), causality is clearly an added quality, projected onto the world, therefore a *secondary quality*. Causal relationships, according to them, can only be understood from a human standpoint. An individual assess the reality as an active agent (even free active agent), and perceives some links as causal. Therefore the truth values of such statements are to some extent dependent on the agent, or his belief system.

For Woodward the matter seems to be more complex, as from one angle he holds that his theory is robust and objective, but does not clearly state that causality is something existing outside of the agent. At the same time he argues that this human-dependence of causality claims is only true for psychological statements. When it comes to facts about the nature, the counterfactuals no longer depend on the belief system of anybody, and their truth value seems independent from the mind of the agent. In such case, the causal relationships exist independently from the agent, our interests as agents serve only as a basis for choosing the kind of relationship between two variables we are interested in, and to evaluate if the relationship is causal. As such the relationship exists independently from our perception of it.

The main feature of the difference making account, an intervention, as I discussed above does not require actual manipulation to occur. Therefore, even if no manipulation is possible, in principle, since we only need to be able to conceptualize it, it could lead us to conclude that there is something as causality, in the world. The theoretical manipulations are then a tool allowing us to discover the causal connections that exist. However, as I argued above, it is not completely clear if these are Woodward's intuitions. Unlike for

Menzies and Price, for Woodward causation actually is an objective fact of the world, independent from the agent's perception, and the manipulation can be used in order to distinguish mere correlations with correlations that are causal and can be exploited through manipulation. However, this only makes causation objective, while not necessarily 'in the world'. It is independent from human abilities, about which the counterfactual claims concerning the result of some, possibly hypothetical, experiment, are true or false. The truth value of those claims is then used to distinguish between the relationships that are, or are not, causal. Yet, this only suggests that causality is in the world, without conclusively confirming it.

Formalization of interventionism.

The definitions of both causes and interventions were only given by Woodward in natural language, without formalizing them, which led to a number of ambiguities. Those depend on the necessity of the two core conditions, manipulability and fixability, and their interpretation can have major consequences to the identification of the causal link.

To quickly recap: manipulation requires that it is possible to construct an experiment, empirical or theoretical, in which the potential causes can be manipulated. If the manipulation of the cause results in a change to the effect, we can argue that the two are connected through a causal link. Fixability, on the other hand, requires that all other variables in our chosen set are held fixed at a stable value, or probability distribution. This is crucial since not fixing other, especially off-path, variables, can lead to false attribution of causal links. Interventions do not have to be performed neither by an agent, not empirically, rather it has to be possible to construct a theoretical thought experiment, where the two conditions will hold.

Depending on how we read the definitions provided by Woodward, the conditions can be seen as joint, or disjunctive, influencing the interpretation of the causal links. Baumgartner (2009) has identifies at least three possible readings of the definition of a direct cause where:

1. *There possibly exists an i such that if i is an intervention on X with respect to Y and all other variables are held fixed, then Y changes its value or its probability distribution.*
2. *There possibly exists an i such that i is an intervention on X with respect to Y and if all other variables are held fixed then Y changes its value or its probability distribution.*
3. *There possibly exists an i such that i is an intervention on X with respect to Y and if all other variables are held fixed and Y changes its value or its probability distribution. (Baumgartner, 2009, pp.165-166)*

Those different formulations are then represented, by Baumgartner, as three different logical formulas, however those differ in two articles written by Baumgartner. In the his formulas C_{vXY} stands for X is a direct cause of Y with respect to a variable set V, where C_v represents the relation of being a direct cause. The predicates F and G stand for being held fixed, and being accompanied by changes in respectively. In the case of the first three formulas G is replaced by H, and there H stands for changing its value or its probability distribution.

In *Interventionism, Causal Exclusion and Nonreductive Physicalism* the formulas are as follows:

$$\forall x, y, z(C_V xy \iff \Diamond \exists i(Iixy \wedge (z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Hy))$$

$$\forall x, y, z(C_V xy \iff \Diamond \exists i(Iixy \wedge ((z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Hy)))$$

$$\forall x, y, z(C_V xy \iff \Diamond \exists i(Iixy \wedge (z \neq x \wedge z \neq y \rightarrow Fz) \wedge Hy))$$

while in the *Logical Form of Interventionism* (2012) the formulas are presented as follows:

$$\forall x, y(C_V xy \iff \Diamond \exists i(Iixy \wedge \forall z(z \neq i \wedge z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Giy))$$

$$\forall x, y(C_V xy \iff \Diamond \exists i(Iixy \wedge (\forall z(z \neq i \wedge z \neq x \wedge z \neq y \rightarrow Fz) \rightarrow Giy)))$$

$$\forall x, y(C_V xy \iff \Diamond \exists i(Iixy \wedge \forall z(z \neq i \wedge z \neq x \wedge z \neq y \rightarrow Fz) \wedge Giy))$$

The two variations of the three formulas different in several aspects, and it is further unclear which one of them is treated by Baumgartner as the correct version, as he does not justify the change between the earlier version, found in *Causal Exclusion and Non-reductive physicalism* to the later version in *Logical form of interventionism.*

The main difference between the two formalizations lies in the placement of the quantifier over z. In the first version z quantifies over the whole expression, which entails that for all z variables in the variable set, x causes y if and only if there is an intervention where we can manipulate x, while keeping z fixed. This manipulation will then result in a change in y. As a consequence it is not necessary to keep all the other variables in the set fixed at the same time, but rather fixing each of the variables individually is sufficient. This however seems to be at odds with both the informal formulation presented by Baumgartner, as well as ultimately with how Woodward seems to conceive of both interventions and causality. Therefore in this dissertation I will assume that the formulation from the Logical Form of Interventionism is the correct one.

When applied to concrete examples the three formulations yield significantly different results. The last reading is the strongest, and Baumgartner initially thinks that it reflects Woodward's position best (as we will see further, this changes).

The first reading allows for the causal interpretations of relation between two variables, when either an intervention connecting the two of them is impossible, or when other variables in the structure cannot be held fixed. We could then therefore identify the connection between two variables as causal, although we cannot exploit it, which would go directly against Woodward's intuitions. As I have already discussed, Woodward thinks that causation is in fact nothing over and above manipulation and the only way to identify such structures is through interventions. This could potentially lead to immediate negation of the presumed causal nature of the relationship between the two events, however, based on other writings of Woodward, it can also result in the case to be deemed ill-constructed.

According to the second interpretation intervention is a necessary condition for the existence of causal connection between two variables, which is closed to what Woodward intended. The problem with this formula lies elsewhere, but the fixability condition is not necessary. This, as I argued before, could lead to unwanted consequence, as false determinations of causal connections.

The third interpretation of the definitions assumes that both manipulability as well as fixability are necessary, to determine if the link holding between the two variables is causal. This reading has been determined to be the closest to, and the best reflection of, Woodward's intuitions and original plans.

What if an intervention is not possible? Are manipulability and fixability necessary conditions?

The identification of the causal, and the non-causal links, on Woodward's account seems to ultimately depend on the interpretation of the two conditions, manipulability and fixability, and what happens when one of them is not possible. Or, in other words, if we cannot fix, or manipulate, can we still say something about the link between the various events? Are such causal claims falsified or just ill constructed? Woodward partially covers this problem in *Making Things Happen*, where he discusses ambiguous causal links, however it is not completely clear what happens when either one or the other (manipulation or fixing the variables) are impossible.

In the absence of possible intervention, for which all the variables involved can be manipulated or held fixed independently from one another, can we still say something about the causal link that can nonetheless connect them? The answer to this question is especially important in the case of mental causation, which will be covered further on in this dissertation. There independent manipulation of the two possible causes, is impossible for metaphysical reasons, the two are connected through a non-causal dependence, supervenience, which forces them to co-change. In such cases, can we still use interventionism to show that a causal link between one of the causes and the effect exists? I have (in my work with P.Verdee, at the time of writing of this dissertation, unpublished) identified at least five different ways to answer this question.

We can adopt the standard approach, where, counterfactuals are true if their antecedent is impossible. Given this interpretation, even if the manipulation, while simultaneously holding other variables in the set fixed, is impossible, the causal claims can still come out as true. This approach stems from a straightforward generalization of the semantics for the material conditional, and is in line with how Lewis (1973a) evaluated counterfactuals with impossible antecedents. Again, this interpretation is likely, also because difference-making is essentially a type of counterfactual approach to causality. This also seems to be supported by Woodward's ideas on mental causation (where the mental still can be interpreted as a cause, while we cannot manipulate it independently from the physical). There are two problems with this interpretation: (1) at least in the case of the mental: it seems a priori, since the conclusions rely heavily on what we choose to manipulate (since we cannot manipulate independently), and (2) does not provide us with an explanation that's broadly applicable.

In the second potential approach the impossibility of independent manipulation, and fixing the variables, falsifies the causal connection. Here the counterfactual is false when the antecedent is impossible, which means that, if manipulation or fixability are impossible, the causal claims are false. Or, they will be false if either one of the conditions is not possible, so not only when the manipulation is impossible, but also when the variables cannot be fixed independently from one another. Since Baumgartner's formalizations imply that both manipulation and fixability are necessary for conditions for an event being a direct cause, we can conclude that the absence of the two results in the negation of the causal link. However, I do not think that this is necessarily Baumgartner's interpretation.

Third position we might adopt is the agnostic approach, where the truth value of the counterfactual with an impossible antecedent, simply cannot be known. I think that, based on his writing, there are good reasons to think this would be Baumgartner's position. If either the manipulation, or fixing other variables in the set, are impossible the causal link holding between the two variables is not falsified, but rather it cannot be determined if the two are connected as a cause and effect or not: we cannot decide either way if the relationship holds or not. Again this approach, when applied to concrete examples, provides us with very little insight about the relationship that holds between two variables.

Another possibility is to deny a truth value to counterfactuals where the antecedents are impossible. When an intervention is impossible, the causal statement is deemed too be lacking the truth value. It differs from the agnostic approach, as in the latter we decide that we cannot determine the truth value of such statement, while in the current approach the truth value simply does not exist. The problem is then not what we can know about causal relationships, given manipulability, but rather what is the nature of such relationships.

The last possibility is a counterpossible approach, where, in order to evaluate counterfactuals with impossible antecedents, we analyze them in such possible worlds, where the antecedent is forcefully made true, at the cost of breaking some laws (laws of nature, metaphysics or logic). Therefore, although the antecedent is impossible, we force it to become possible. This approach might also be in line with what Woodward wants to achieve, especially in connection to the solution of the Causal Exclusion problem (as we will see in the upcoming chapter). This option has the added advantage of allowing us to still determine the, at least theoretical, truth value of the statement, although manipulation is impossible because of some nomological reason. Other reasons to believe that counterpossible approach is the best fit, is Woodward's rejection of not only the claim that an agent has to manipulate the cause, but also because he seems to clearly state that we do not necessarily have to know the processes that take part in the causal relationship (discussed previously in this chapter). It would also be in line with his apparent thoughts on the cases where experiments are impossible (and we should still think of scenarios where they would be possible anyway).

Given all these possibilities, it seems that also in cases where either independent manipulation, or fixability are not possible, not all is lost for interventionism. In fact, only in one case the conditions of causality are strict enough to imply that in cases where manipulation and/ or fixability are impossible, the connection between two variables is falsified.

Causation in neuroscience.

While in philosophy the notion of causality is one of the major puzzles yet to be solved, surprisingly little is being said on the topic in neuroscience. Being an experimental domain, with the core assumption of experiments being manipulation, it seems that interventionism, or difference making, would be the more appropriate approach. Yet using interventions is different than committing ourselves to the view that causality is nothing over and above making a difference (as Woodward does), and the latter might not necessarily fit with neuroscientific intuitions.

Studies using manipulation (e.g. transcranial magnetic stimulation) aim to manipulate the electrical activity in the brain, and it might be the case that the production account of causation would be a better fit, since electrical and chemical currents are exchangeable physical quantitates (e.g. Sack, 2006), and would be covered by the types of quantities permitted by Salmon and Dowe. Since, in brain stimulation studies, usually only the relationship between various brain states is explored, and not between brain states and mental states, it could suggest using two different approaches to causation depending on the application. Few authors discuss their notion of causality directly (e.g. Dijkstra, de Bruin, 2016), and they seem to lean towards the difference making approach.

Both production and difference making accounts have their advantages and difficulties, when applied to discussions on human action. At the heart of production accounts is the physical connection between the cause and effect describing how the effect is "produced" by the cause. In neuroscience this type of causation could be used to describe the causal relationship of the process of the signal being passed between neurons. However, this exact process was used by Craver to argue that the mechanistic approach would serve best, when applied to neuroscientific research. Moreover, the process accounts, while highly applicable in physics, face issues when applied to sciences such as biology (Machamer et al. 2000) or social science (Russo 2009). Assuming that there are causal processes beyond the level of basic physics, in both fields the processes and interactions that are observed and described as causal, are impossible to describe in terms of transmission.

If we look again at approaches in line with Libet's methodology, the influence readiness potential has on the physical action can be described by both: processes and mechanisms. But, the same cannot be so easily achieved when we try to do the same for the causal influence of the conscious decision. Here, again Craver's account could be applicable, as he assumes that multi-level explanations are a necessity in neuroscience, as explanations given only in terms of transference of physical properties might just prove to not be robust enough. As we will see further, explanations involving different levels are deeply problematic, and require explanations of the dynamics between them.

Interventionism exploits counterfactual statistical dependencies, in order to investigate the nature of a relationship between two variables, which, again, seems highly applicable, especially to approaches such as Libet's. Yet, it is unclear if full blown interventionism would be acceptable for experimental scientists.

Conclusion of chapters 4 and 5.

In the first three chapters of this dissertation I have explored both the classical debates on the free will problems, and the ones linked to the experimental results provided by neuroscience. I have argued that the problem of causality is present in all of them: determinism is seen as causal determinism, and the experiments investigate the causal source of the action. This led me to conclude that, in order to proceed, we need to understand what do we mean when we talk about causality.

In these chapters I have discussed both the classical notions of causality, such as Hume's, Anscombe's, Davidson's and Lewis, and their modifications, as Mackie's. Then I have proceeded to the modern approaches, focusing on the production accounts (processes and mechanisms) and difference-making accounts.

I have shown that, when applied to neuroscience, the three modern positions all have their advantages and disadvantages, however in order to proceed in discussing the connection between causality and free will, it is paramount to identify what kind of causes are taken to be necessary in the generation, or production, of intentional action.

Most important points from the chapters 4 and 5 (in a few short points):

(1) The classical approaches are law-centered, where causality is somehow backed up by the laws of nature. For Hume those laws are inaccessible for the agents. What we are able to perceive are only co-occurrences, which might be coincidental, we don't have a definitive way to say either way.

(2) Most of the discussions, also the modern ones, are focused on finding one, necessary and sufficient cause for an occurrence. Mackie argues that this might not be possible, since all causation happens in a context, and that context might play a role in causation. Instead Mackie proposes INUS conditions, evaluated with the background conditions, the causal field, in mind. This position, as we will see further, might have consequences for the free will debate.

(3) Anscombe pointed out that the notion of causality is very vague, and there are more specific descriptions, that are better suited and more informative, than the umbrella term. She also pointed out the language dependence of the causal links. This point she shares with Davidson, for who different causes are only different in the way they are described.

(4) The production accounts have departed from the usual law-backed approach to causality, understanding the links between the two events as causes producing effects. Process accounts, mainly transfer theories, involve an exchange of a physical quantity, and have the potential to explain some of the neuroscientific findings. However their applicability outside physics is limited, and I have questioned if they would serve well to attempt to answer the types of questions asked in the experiments on free will.

(5) Mechanistic approaches, while developed (at least by Craver, 2007) with neuroscience in mind, and using examples of the brain processes, when applied to the question of free will, as we will further might suffer the same issues as the process accounts.

(6) Difference-making accounts aim to capture the intuitions of the scientific experiments, where the cause is taken to make a difference to the occurrence, or the probability distribution of the effect.

The next chapter, on the causes of our actions, will focus on analyzing the way the action has to be caused, to be considered free or voluntary. This will influence the applicability of the different approaches to causality.

Chapter Six. What are the causes of free action?

In the previous chapters I have summarized the main discussions on the free will problem (chapter 1 and 2), focusing mainly on the compatibilism of free will with determinism, and the principle of alternate possibilities. The positions regarding the former issue can be divided into three main categories, illustrated in the table below.

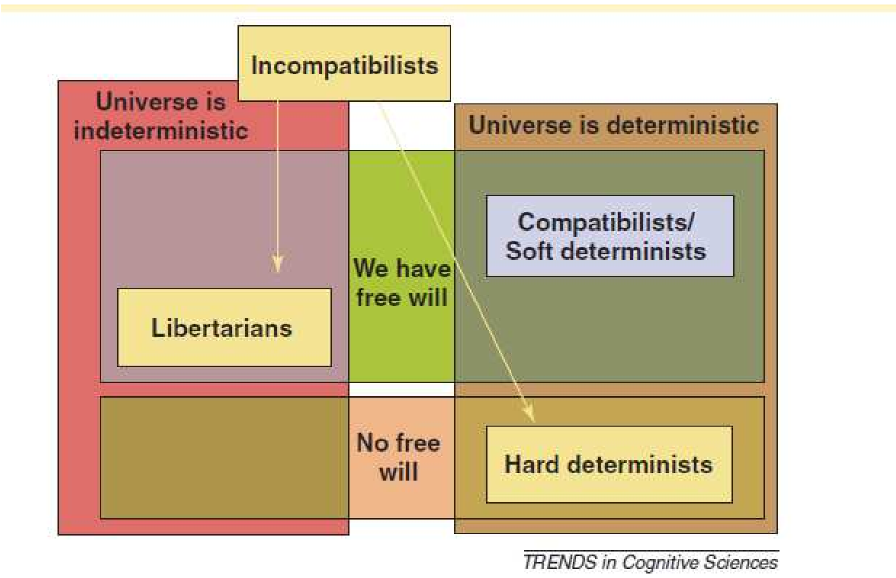


Figure 8. Compatibility of free will and determinism. Source: Roskies, A. (2008)

Compatibilists argue that, even if the universe is deterministic, often not decisively claiming that it is the case, free will is possible. Libertarians on the other hand argue that free will is inherently incompatible with determinism, and the universe is indeterministic (often quoting Quantum mechanics as a reason for this conclusion). Hard determinists, as libertarians, argue that free will is not compatible with determinism, and moreover, potentially the concept of free will itself is incoherent (the latter is not accounted for in Roskie's diagram).

In the second chapter I have focused on the problem of principle of alternate possibilities, and the question whether the different options, for the individual to choose from, have to exist. I have discussed Frankfurt's objection to the claim, where he presents a revised version of the principle, stating that the actual options are not as important as the reasons for which the action was performed. In case the agent did something because he had reasons to do so, regardless of the nature of the universe, his action can be considered free.

In the third chapter I have analyzed the other side of the discussions on free will: the neuroscientific experiments. I have used the famous Libet's experiment as a blueprint of the discussion, as, since its original run, it has been replicated and modified a number of times. The illustration below summarized the experiment:

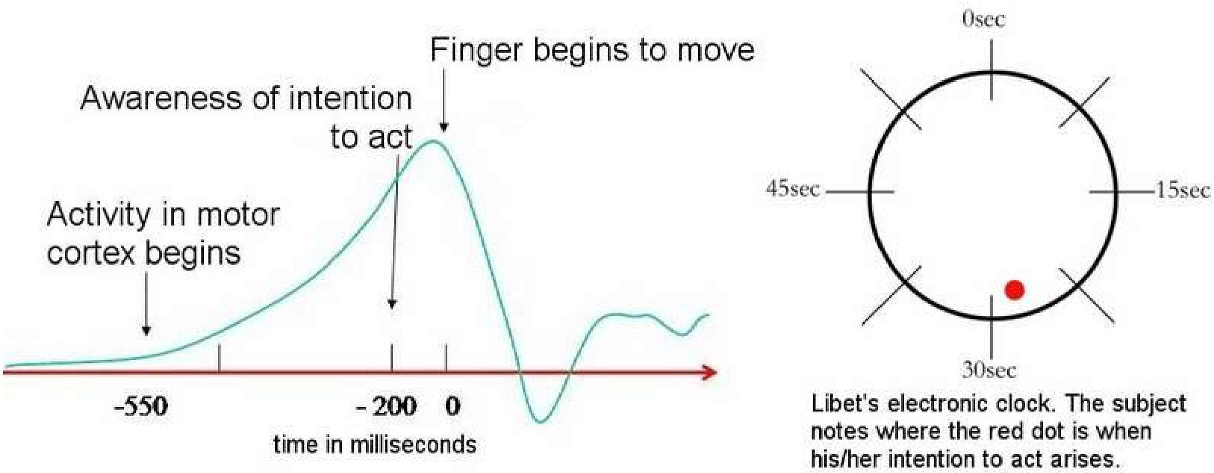


Figure 9. Results Libet's experiment http://www.world-of-lucid-dreaming.com/is-free-will-an-illusion.html

The experiment showed that a consistent spike of activity preceded the overt movement by, on average, 550ms, while the conscious decision was prior to the movement by, on average, 200 ms. This led Libet to conclude that, since the action is not initiated consciously, free will as we know it is impossible, however it can be preserved as a veto power over actions already underway. In the chapter I have provided ample evidence that Libet's argument might not support the conclusions that either free will does not exists, or our actions are not generated in a way that would support free will, and it has to be limited in its scope to influencing actions already underway.

The discussions presented in the first chapters emphasized the importance of causal links for the debates on free will. Firstly, the causal determinacy, which seemingly, although not necessarily, entails the predictability of all future, leaving little room for free action. For Frankfurt, on the other hand, causality is determining what sort of causes are at play in action generation: if it is only the antecedent conditions combined with laws of nature that cause the action, it cannot be considered free, whereas when they are accompanied by reasons, it can be. However, unlike neuroscientists, Frankfurt does not seem concerned with the actual causality of the reasons, and it seems that just their mere existence is sufficient for the action to have the potential to be free.

The experimental approach seems closed to Frankfurt's position than to classical debates on the topic. However, the problem here seems to be twofold, on the one hand, we have again the problem of predictability of action, based on interpretations of the experimental results by, e.g., Soon et al. If the action cannot be completely predicted from the physical brain activations, it seems to leave little room for free will. On the other hand, the question of the causal connection comes more into the foreground here, as two potential causes are seen as competing with one another: the physical brain activation and the conscious decision. If the action is caused by the former, it is not free, if it is the latter, it can be. Two problems emerge here: (1) what are the types of causes that can generate the free action, and (2) what kind of notion of causality we are operating with. The three problems: determinism, predictability and causality are often discussed as interconnected. However, I used the first chapters to establish that different problems can come from each of them independently. The next chapter will serve to further establish the differences between them, but, so far they can be summarized as follows:

(1) Determinism- is free will compatible with determinism, or requires indeterminism. Determinism is clearly understood as causal determinism, without clarifying which understanding of causality is applied. Most likely is the humean approach, where the observable co-occurrences are backed by laws of nature. This understanding would also fit with the next notion, or predictability. The problems with determinism have to do, mainly with the lack of choice, a feature that seems essential for a truly free action.
(2) Predictability- it is seen as a threat to free will for, what seem to be, two different reasons: (1) the deterministic nature of the universe in laplacean understanding- is there is a being that knows everything about all the antecedent states, and laws of nature, can, in principle, predict the whole future, and (2) our actions are predictable based on the brain activations, which is how it is seen in neuroscience. The latter

understanding can be seen as much less threatening to the notion of free will, also it is much less metaphysically heavy.

(3) Causality- not discussed independently from the previous two points. The consequences of the types of understandings of causality to the free will debate are often not discussed. This chapter will serve as providing a more solid basis to the claim that causality, taken independently from two previous problems, causes serious problems for the possibility of free will.

The question of what causality means, has been further explored in the chapters 4 and 5. Since the notion used in the neuroscientific experiments is unclear, I have offered a number of possible solutions, which could have the potential to do the work seemingly required by neuroscientists. I started with the law-based solutions, such as: (1) the regularity theory by Hume, and its modified versions by Mackie and Davidson, where causality is taken to be observable co-occurrences, supported by laws, and (2) counterfactual approach of Lewis, where causality is synonymous with counterfactual dependencies. I have also explored contemporary accounts, which, instead of focusing on the elusive notion of a law, focus on the types of connections that are needed for the cause- effect relationship, such as: (1) production accounts, where the cause is seen to produce the effect, either by means of transfer of a quantity, or through mechanisms, and (2) interventionism, where the cause is taken to make a difference to the occurrence of the effect.

Current chapter will focus on what kind of causes can be seen to generate, or produce, the type of actions we are identifying as free. In the discussions on free will in neuroscience, almost exclusively the cause of an action is taken to be the agent's conscious decision, however the notion is not clarified, nor properly defined relying on intuitive understanding of it (and it is often mixed up with notions such as urge, desire etc.). The conscious decision does not exhaust the possibilities, as for some (e.g. Haggard) other potential causes, or causal contributors, are, for instance, intentions.

While, as we have seen in the first two chapters, free will is usually discussed as a metaphysical issue, the question of voluntary action, and its causes is usually the focal point of philosophy of action, rather than metaphysics. Therefore, most of the positions discussed here, will have little to say on the question of compatibility of these causes with determinism, or predictability of action. To remain as clear as possible, I will again treat the terms voluntary, intentional, and free as synonymous. While in other contexts this might not be the case, I think that every free action, in the most intuitive sense of this word, has to be both intentional and voluntary, or at least voluntary. There might be much more to say on this topic, but, to remain focus, I will assume that the actions I, and the neuroscientists, would like to explore, are both intentional and voluntary, therefore the terms will be used interchangeably.

As was initially established, in the experimental approach, for an action to be treated as voluntary, it has to be caused by a conscious decision, as opposite to an unconscious brain activation (e.g. Libet, 1983, 1985). In philosophy of action the possibilities are however much broader, ranging from desires and beliefs (e.g. Davidson, 1980), through reasons (e.g. Habermas, 2008) to intentions (Mele, 2009), and mental images (Nanay, 2016). I will begin this chapter by analyzing the different options, and I will argue that the thread connecting them all is their mental character. If this is the case, the problem of free will is connected not only to the notion of causality in general, but also, or maybe mainly, to the more specific problem of mental causation.

Similarly to the discussions in the experimental domain, also here it is not always clear what kind of a notion of causality is being used, but some traits of it are applied generally throughout the discussions. Those include the diachronic character of the causal relationship, where the cause is taken to precede the effect. Another angle of the discussion, and a source of ambiguity, is the type of things that enter in those causal relationships. The standard view, of the event-causal theory, takes the relata of such relationships to be events (e.g. Davidson, 1980a), but some argue that wither facts (e.g. Bennett, 1988), features (Dretske, 1977) or situations (Menzies, 1989) would be more appropriate. While the event causal theory has its problems, in the specific case of free will the disappearing agent, I will treat the two as events, since this is the dominant approach.

I will start the chapter by discussing the two authors I have already mentioned in the chapter on causality, namely Anscomve and Davidson. Both of them create a complete account of, not only what should be considered causal in voluntary action generation, but also what is causality. They also represent two main positions with regards to causality of reasons: anti-causalism and causalism respectively, which can be defined as follows:

Causalists hold that all explanation of action presupposes the existence of strict causal laws. Conversely, anti-causalits commit themselves to the view that action explanation is not causal in the relevant sense. (Sandis, 2012, p.46)

Which is then illustrated in the table below:

	Claims counting for (+)	**Claims counting against (-)**
Causalism	The events of our doing must be explained causally	The explanation of why we must act as we do must cite a cause
Anti-Causalism	The explanation of why we act as we do need not (and typically does cannot) cite a cause.	The events of our doing things cannot be explained causally.

Table 1. Causalism and anti-causalim (Sandis, 2012, p.47)

Reasons, taken to be causal by Davison, are non-causal on Anscombe's view. The core of their debate is, however, not the type, or the nature of the causes of voluntary actions, but rather if the causal explanations cite actual causes or not, where:

According to causalists about action explanation, one necessary condition is that the explanation cites a cause of the action. Anti-causalists about action explanation disagree. (Mele, 2013, p.160)

As was already discussed here, causal explanation and causation are not necessarily the same, yet they are related. In the previous chapters both sides of the discussion were analysed, focusing on both the type of explanation that can be applied to the context of intentional action, and on the causes that have to be 'productive' in the generation of action.

The two, essential, issues, although important, make little difference for this chapter, as here the core question is **what** is taken to be causal at al. I will then treat causal explanation of the action, as connected to causality, and quoting at least one of the causes for the action. Therefore the debate on causalism and anti-causalism is relevant for the different types of causes that are being discussed, and less so for its core arguments, namely can reasons be causal at all. In the end, being able to give reasons for an action as a condition of it being voluntary, presumes that said reasons can be truly causal in the action generation. In this sense I am following the way of framing the problem by Davidson, who aside from arguing for the causality of reasons, also provided an account of how mental causation could have causal impact on the physical world (anomalous monism). Approach assumed in this dissertation is based also on the way results of experiments are interpreted, where causal explanation is taken to, at least, cite genuine causal connections.

Throughout the chapter I will take a look at different understandings of the causes of behavior, focusing mainly, as is usually done in the field, on reasons and intentions. I will then move to the assumptions in the neuroscientific domain, where the decisions are taken to be the causal basis for the behavior. I will also refer to issues stemming from the proper definition of an event, process and action. The latter being of particular importance, as some (e.g. Alvarez, 2013) argue that movements are not actions, but rather results of an action. In my own position, taken into account the neuroscientific interests of this thesis, I will treat action, including voluntary action, broadly and I will include simple movements into this category. Since I seek to apply the, well-established in philosophy of action, notions in the specific context of the problem of free will, I will focus on how they could be causal, and ultimately bring the discussion to the problem of mental causation, however I recognize that it stems more from my particular project.

Anti-causalism and the work of G.M. Anscombe.

The causalism/anti-causalism debate has been one of the classical issues in philosophy of action. As we saw in the, rudimentary description above, causalists argue that reasons cite actual causes of our actions, while anti-causalists argue that reasons do not play such role, and propose to replace them with e.g. intention instead.

Most important work on intention in the framework of the anti-causalist debate is *Intention* the work by Anscombe, published in 1957. It also marks a clear turn in the, broadly accepted to this point, causalist approach to how our actions are caused, although Anscombe's approach was heavily influenced by that of Wittgenstein.

The intentions discussed, are divided into at least three types: (1) intention to act; (2) intentions in acting and (3) intentional action, where intention to act is directed to the future, and presumably plays a significant role in the action-generation. A proper account of intention should however encompass all three meanings of the notion, and additionally explain the relationship that holds between them. The inclusion of intention, as one of the potential causes of action, was seen as part of strictly anti-causalists, however as we will see further, already Davidson recognized that excluding intention, or re-describing it in terms of reasons might be problematic. Currently intentions are often used for action explanation purposes by both causalitsts and anti-causalists alike. It is worth remembering, discussed already in the chapter on causality, that Anscombe thinks of the notion of the cause, as far too broad to be truly informative, preferring more specific notions.

In the discussions on intention, Anscombe start from identifying the differences between two, seemingly closely related terms: intention and prediction. While we, as agents, are rarely mistaken if an utterance expresses one or the other, e.g. somebody stating that they are going to be sick is most often seen as a prediction, while when they say they are going for a walk, it is considered to be an expression of an intention. Yet, both of them share many characteristics, including their future orientation, something that seems to be the core of intention understood as a trigger for an action. For intention to be seen as a cause of an action, and because of the diachronic nature of causation, the former has to be about the future, in order to cause the latter, e.g., I go to the sink, because I have an intention to drink water.

Not all cases including either prediction or intention are equally clear though, as sometimes their content might be similar and sentences like: e.g. a thought 'I am going to hit the cat with my car', depending on a broader context might be either one of those. Thinking 'I am going to hit the cat with a car' while I am driving, just a split second before I hit a cat by pure accident, might not signify my intention to hit a cat. Rather is a prediction that could be understood in terms of 'there is nothing I can do in order not to hit the cat'. In a different context, when for example we hate the neighbour's cat and want to kill him, might signify an intention. In that case we would be driving on the street, see the cat and form an intention 'I will hit the cat with the car', thinking that this will finally free us from the hated animal (gruesome as might be). Focusing on the prediction's orientation to the future offers little help, as intention can have a similar orientation. Stating that prediction, unlike intention is future-oriented would simply be false. Anscombe even states that, not accidentally, the intentions are always concerned with the future events, since

No one is likely to believe that it is an accident, a mere fact of psychology, that those states of mind which are intentions always have to do with the future […]"(Anscombe, 2000, p.2)

Here, again, the mental character of the causes of action is emphasized, as intentions are called states of mind. Therefore providing both: an account of how the two, the mental and the physical, are related, and an account how the mental can be causal, seems inevitable.

Not only intentions and predictions but also intentions and motives seem to be closely related. Motive however, seems to be a wider notion than that of intention and, moreover, motives cannot be considered causes (similarly to reasons on Anscombe's account).

In Anscombe's typology, another interesting feature is the possibility to differentiate between actions that are intentional and actions done because of a specific intention. Similar intuitions can be found in the classic Knobe effect. For Anscombe we can differentiate between situations, where

5) The agent intends to G;

(6) The agent G'd intentionally; and

(7) The agent F'd with the intention of Ging,

For that matter, one could add

(7') In Fing (by Fing), the agent intended to G (Winston, Shpall, 2016)

Since the differences between prediction and intention are not completely clear, and are difficult to conclusively identify , especially when the same sentence, like the ‘I am going to hit the cat with the car’ can be either a prediction or an intention, making language unreliable tool for differentiation between them, Anscombe focuses on the intentional action instead. Two issues seem to surface here, firstly the question of language. As we will see further on, Davidson’s main tool is the language, and the mental is distinguished from the physical based on the way event is described. Secondly, Anscombe seems to assume that what distinguishes an intentional action from a non-intentional one is more than just a description.

In case of Knobbe, the difference between intention and intentional action is at the core of our analysis of what is morally good, and what is morally wrong, and by extension what we are seen to be morally responsible for. If we consider free will as related to moral responsibility, as most people do, this differentiation might have consequences to the responsibility attribution. Knobe presents two scenarios:

Harm

The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.” The chairman of the board answered, “I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.” They started the new program. Sure enough, the environment was harmed.

Help

The vice-president of the company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.” The chairman of the board answered, “I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.” They started the new program. Sure enough, the environment was helped. (Knobe, 2003, p.191)

In the harm scenario most of the participants said that the chairman was intentionally harming the environments, therefore attributing the blame to him. On the other hand, in cases where the environment was not harmed, the participants did not ascribe this to the intentions of the businessman. Knobe ascribes this difference to the sensitivity of the intentional action, or its interpretation, the problem of side effects of an action.

The switch, from intention as such to intentional action, allows Anscombe to focus on conditions which have to be met if the action is to be identified as intentional. Again, similarly to the contemporary positions, she seems to reject dualism, by recognizing that the mental is somehow connected to the physical. The intention is not something confined to the realm of the mind, but rather has a physical 'anchor', when she argues:

All this conspires to make us think that if we want to know a man's intentions it is into the contents of his mind, and only into these, that we must enquire; and hence, that if we wish to understand what intention is, we must be investigating something whose existence is purely in the sphere of the mind; and that although intention issues in action and the way this happens also presents interesting questions, still what physically takes place, i.e. what a man actually does, is the very last thing we need consider in our enquiry. Whereas I wish to say it is the first. (Anscombe, 2000, p.9).

The existing line of inquiry on intention is described as one that focuses mainly on intention 'in itself', without the connection to intentional action, an approach that Anscombe wants to change. By starting from the behavior, the movement, or what *the men actually does*, Anscombe's approach seems closer to the experiments conducted to investigate free will, and, moreover allows for distinguishing between intentional action, and one done with a specific intention in mind (as in Knobe effect).

As we have seen previously, in case of Libet, the starting point of his investigation is also the movement rather than the, in his case, conscious decision in isolation. Of course this manner is forced by the construction of the experiment, where the movement is necessary to establish a starting point, which serves as an 'anchor' in checking what happened in the brain before it. For Anscombe this way of thinking about intention seems to emphasize not only the existence in the causal relationship, but also to distinguish what types of action follows intentions. The classification of action as intentional entails that, some form of a 'why' question can be applied to it. Therefore an action is seen as intentional, if we can ask 'why' in relation to it. On Anscombe's account mental causation has a broader span than just action causation, but can influence a number of processes:

Mental causes are possible, not only for actions ('The martial music excites me, that is why I walk up and down') but also for feelings and even thoughts. In considering actions, it is important to distinguish between mental causes and motives; in considering feelings, such as fear and anger, it is important to distinguish between mental causes and objects of feeling (Anscombe, 2000, p.16).

Again motives are mentioned, and as previously they were distinguished from intentions, here they are also set apart from mental events as such. Anscombe seems to recognize not only mental to physical causation in the passage quoted above, but also mental to mental causation, where mental causes cause mental effects such as emotions.

As I have previously stated, in this dissertation, because of its specific focus on free will, I am inclined to treat the intentional and voluntary as synonymous, but it has to be noted this is not the case for Anscombe. On her account the two: voluntary and intentional, can be further distinguished, although it seems that most of the time they fall in line. The situations in which they diverge, and the action is voluntary rather than intentional, include ones where the answer to the 'why' question is something along the lines of 'I don't know', or the movements executed by the individual are not considered by the agent. Something is also voluntary but not intentional when

[…] it is the antecedently known concomitant result of one's intentional action (Anscombe 1957, p.89),

or when it happens not for one's doing but the agent consents to it. Actions can also be intentional but not voluntary, when the agent regrets doing them. It additionally seems that actions such as ones in Knobe effect, discussed above, could be another example when doing something voluntarily and not intentionally is possible: while the business men decision to preserve the environment was voluntary, in the sense he was not coerced into doing, it was not intentional, in the sense he did not have an intention to preserve the environment. It also seems that the two diverge mainly in cases where we cannot clearly identify the cause as mental, e.g. where we do not know why we have executed a movement, it seems we could conclude the movement was caused by some purely physical cause (assuming that those differ and mental can be causal). This typology could be used against my assumption of the simultaneity of the various causes, however, I choose here to focus on the actions that are voluntary, intentional and free, to limit the potential problems (as there are already enough of those).

Anscombe moreover argues that the sole fact reasons usually precede actions is not sufficient to conclude that they cause the action, since other events, e.g. motives, also precede it and yet they are not causes. Similar arguments were used against Libet's experiments, with some (e.g. Schlosser, 2014) arguing that Libet only establishes time sequence of events, which is not synonymous with causality of one over the others. The causal relationship is usually taken to be diachronic, with the cause preceeding the effect temporally. However, in the relationship between the two potential causes, the temporal primacy does not have to guarantee causal impact. On Anscobe's account the importance of reasons in identifying action as intentional is, to some extent, preserved, where the action can be identified as intentional if reasons for the action can be given. This does not however entail the conclusion that reasons are causal. However, in spite of its apparent importance, mental causation as such does not seem to be central for Anscombe, and she offers little to explain how mental intention, could be causal in action generation, given the physical nature of the world.

Another, related problem is defining what action actually is, and how can it be qualified as intentional. This problem was already briefly introduced in the beginning of this chapter, in the argument by Alvarez (2013), who thinks that muscular activity does not itself constitute an action, but rather is a result of an action. Her position is in fact based on that of Anscombe, who argues that:

In describing intentional action as such, it will be a mistake to look for the fundamental description of what occurs- such as the movements of muscles or molecules- and then think of intention as something, perhaps very complicated, which qualifies this. The only events to consider are intentional actions themselves, and to call an action intentional is to say it is intentional under some description we give (or could give) of it. (Anscombe, 1957, paragrah 19, p.29)

Here, just like on Alvarez account the action is not a physical, muscular movement, qualified as intentional because of how it was caused, rather the movement is a part of an intentional action, along with the intention. This approach seems radically different from the one seemingly used in the neuroscientific experiments on free will and voluntary action, where the movement is taken to be an action in itself.

Anscombe's approach was described by D'Oro and Sandis as a:

[…] a rejection of both the causal theory of action (which claims that an action is a bodily movement with a special sort of cause) and certain versions of the causal theory of action explanation, namely those which claim that we can only ever explain an action by pointing- be it directly or otherwise- to some internal, private mental cause. (D'Oro and Sandis, 2013, p.18)

The action is then no longer described as a movement, or another physical event involving muscular contractions, and because of it, the explanation of action also has to be modified, and can no longer just quote just a mental, or I presume a physical, cause of such event. As a consequence, mental causation itself seems pushed to the foreground, as the primary issue seems to be proper identification of what an action is, and what are its potential causes.

In short, Anscombe rejects then not only the old consensus, that an intentional action is a bodily movement, same as any other movement, different from its nonintentional counterparts because of the way it was caused. She moves away from the manner of discussion where the action itself is of less importance, where the main problem lies with what causes the action. As we have seen a similar approach is adopted also by Alvarez (2013), who argues that movements are not actions, but rather results of an action.

Following Anscombe's approach, also the results of the empirical research on free will, like Libet's, would be invalidated, since it goes in line with the classical approach where the movement constitutes an action, and has two potential, somehow possibly independent, causes. Some, like Haggard (2011) equate the problem of free will, as the problem of causes of our overt actions, and Anscombe's argument could justify a deeper investigation into what actions and their causes actually are. This also furthers the doubts about the neuroscientific approach into investigating intentional, voluntary and free action, and the possible causes of it.

The classic belief/desire models of intentional action: reasons as causes.

In the classical debate on how the intentional action or intentional behavior comes about, the role played by beliefs and desires of an agent, or the reasons for actions is emphasized. In this standard model, the reason explanation is sought, where

[…] *an intentional action A supposedly explains the agent's A-ing by identifying the reason(s) for which he A-ed.* (Mele, 1992, p.122).

The standard model has been a part of the general consensus from the times of Aristotle, Hobbes and Descartes. The model at its center has the assumption that human action is an outcome of causal relation that takes place between some form of mental event, and a movement, and as such was opposed by anti-causalists. Causalism, as the theory propagated by Davidson, can be defined as:

[…] a theory of what it is to behave intentionally, is the view that an agent behaves intentionally when certain events/states involving the agent's body (such as the agent's moving in a certain way) are appropriately caused (non-deviantly caused, or caused in the "normal" way) by certain mental events or states of the agent, in particular, the agent's intentions, belief and desire pairs, decisions, etc. (Santorio, 2009, p.514)

Therefore, opposite to Anscombe's position, a movement constitutes an action, and it can be seen as intentional (and, for this dissertation also as free), when it is caused in a specific way. Davidson's position is also in line with how the question is experimentally investigated, and is a part of the so-called Causal Theory of Action, three essential properties regarding human action are :

(1) such actions are (at least partially) identical to movements of the human body,
(2) those movements are done in response to certain representational states of belief, desire, intention, volition, willing, choice, deliberation, or the like, and
(3) the 'doing' in (2) is analyzed in causal terms (Moore, 2010)

Davidson's point of departure into the 'right way' in which a voluntary action should be is the notion of rationalization. He argues that the usual rationalization of action, by giving reasons for which action was executed, is a species of a causal explanation. The reasons should therefore be considered causes of an action. Having a reason on the other hand, is described as containing two, distinguishable, components:

[…](a) having some sort of pro attitude toward actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind. Under (a) are to be included desires, wanting, urges, promptings, and a great variety of moral views, aestethic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agnet directed towar actions of a certain kind. (Davidson, 1980a, p.4)

Said reasons are then seen as a combination of agent's belief that something is the case, and his desire to reach a specific goal. This can be illustrated as follows:

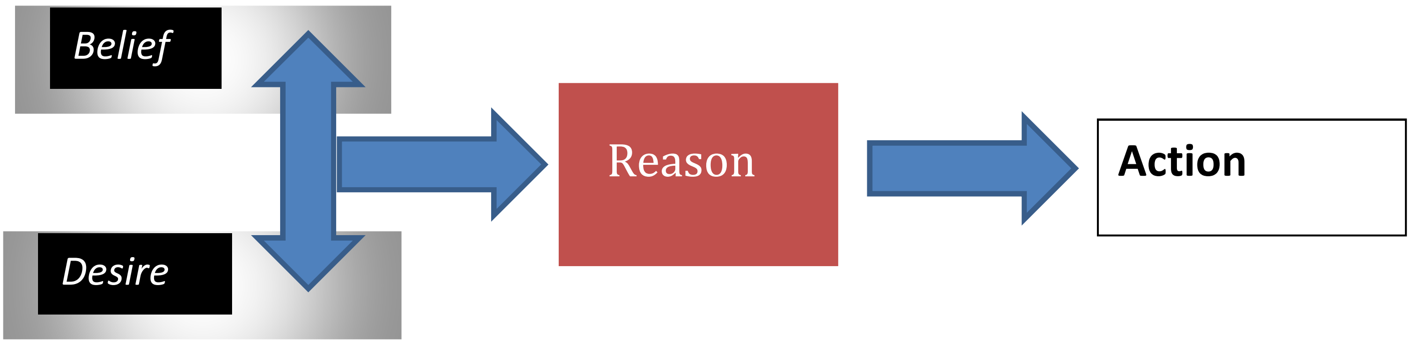


Figure 10. Davidson's model of action generation.

This can be illustrated with a simple example, where my reaching for a beer in the fridge on a sunny day is caused by the combination of my desire to drink a beer and my belief that there is beer in the fridge. Therefore I can give reasons for performing this specific action, of reaching to the fridge to get a beer. This pair: belief and desire, together constitute a primary reason for action, which, for Davidson, is synonymous with its cause.

In cases, which seem to be frequent, where the agent causes more than one effect (here Davidson gives an example of a person switching a light switch and besides illuminating the room, also alerts the burglar), the primary reason explains which of the effects was in fact intentional. Importantly, it seems that each of the effects, on Davidson's account constitutes an independent action, rather than the same action under different description. To know the primary reason is to know the intention that guided the action execution. Although initially Davidson argues that:

There are no intentions. When we act intentionally, we act with an intention, but the locution 'with the intention of' is syncategorematic. (LePore and McLaughlin, 1985, p.6)

Later on he changes his mind, and includes intentions as irreducible to reasons, and causal in action generation processes.

Davidson argues that, since the so-called rationalizing explanations, a set that includes primary reasons for acting, are normally used in action explanations, this would entail that the primary reasons are in fact causes of actions. Significantly, for Davidson, there is a connection between causation and causal explanation, and the latter quotes actual causes for an action or behavior. According to him, the anti-causalists fail to explain how the reasons could be used to explain the action, but at the same time would not cause the action in question. He also recognizes the complexities stemming from the ambiguity of our language, where the same action can be re-described in a number of ways, placing it in a pattern, but for him such pattern is causal.

The core ideas that form Davidson's causal theory of action can then be summarized in four premises:

(1) In order to understand how a reason of any kind rationalized an action, it is necessary and sufficient that we see, at least in 'essential' outline, how to construct a primary reason (Davidson, 1980a, p.4)
(2) The primary reason for an action is its cause (Davidson, 1980a, p.4)
(3) R is a primary reason why an agent performed the action A, under description d, only if R consists of a pro-attitude of the agent towards actions with a certain property, and a belief of the agent that A, under the description d, has that property. (Davidson, 1980a, p.5)
(4) A rationalization that cites a primary reason is a species of causal explanation (LePore, McLaughlin, 1985, p.5)

The premises, while forming a strong basis, are not exhaustive, and moreover, do not form necessary and sufficient conditions for reason-causal action explanation, an issue Davidson himself recognized, and argued that more steps need to be added to achieve that.

The additional assumptions include the, already mentioned, rejection of intentions as causal, as well as the possibility to deductively relate the propositional content of the primary reason, to the proposition on the desirable characteristics of the action. For example: we add spices to our food in order to make it taste good, but also adding spices is desirable because it improves the taste of the food. Moreover, Davidson also maintains that every action, under some description, has at least one primary rationalizing explanation.

Davidson's reasoning can be, then, summarized as follows: (1) the primary rationalizing explanation explains an action through the primary reason, which was the cause of the action; (2) said primary reason is constructed by two mental events: a belief and a pro-attitude towards the action; (3) primary reason is a reason for action because of the existence of a logical relationship between the propositional content of the belief and the pro attitude combined with some desirable characteristic that the action possess; (4) there is a causal connection between the pro attitude, belief and an action, therefore the explanation that cites the primary reason is a causal explanation (following LePore and McLaughlin, 1985, pp.9-11).

As for the discussion on free will, and the types of causes that are relevant in the free action generation, both reasons and intentions (after he accepts them as independent from reasons) seem to be mental entities. Although distinguishing them, mainly, based on the type of description used, Davidson goes a step further than Anscombe, and provides an account which explains the causality of the mental in the physical world (anomalous monism, which will be covered further on). This, of course, is closely linked to Davidson's general approach to what action is, where a movement is an example of one, and it can be caused by, presumably, either the mental or the physical event.

This general approach to how actions are caused, and to causal explanation, is closer to both the philosophical discussions as well as the experimental approach in the arguments on free will, than the one presented by Anscombe. Firstly, just like in the various experiments, it supports the thesis that an intentional, or voluntary, action is one that is caused in a specific way. Secondly, it sees a simple movement as a type of action, that could moreover be considered intentional. As such Davidson's approach seems more appropriate when experiments are considered, however the cause of the action differs, for Davidson reasons are the causes, while for neuroscientists it is the conscious decision.

Qualifying reasons as causes can be understood in a number of ways, with one of them having to do with the ability of sciences such as psychology to provide us with genuine causal explanations. The question goes beyond specific action-explanations, touching issues like the similarities between the types of laws described by psychology and ones used in physics, where such laws, combined with the antecedent states, we can predict the behavior of an individual. Here, again, we can see the parallels between the question of what causes our actions, with the issue of causal determinism, essential for the debate on free will. If we see causal determinism the way Laplace did, and the laws in psychology have the same properties as the ones in physics, based on them and the previous state of the universe, the behavior of the individual could in principle be predicted.

From the angle of mental causation, for some (e.g. D'Oro and Sandis, 2013), if, on the causalist account, the beliefs are to retain some of their causal powers and not be epiphenomenal, they have to play a physical role, either directly, or through their physical realizers. The last sentence is particularly important, as, often in the discussions on free will, the problem of mental causation is pushed on the side because it is assumed that it is sufficient for the mental intention, or conscious decision to be causal either by itself, or through its neural realizers (e.g. Mele, 2014). This argument however seems to, at least at first glance, assume that at least sometimes it is the conscious decision or intention that is causing something by itself, and not through its neural realizers. This, in turn, would still require that the mental component have the potential to be causal, and therefore requires an account of efficacious mental. The problem seems to lie not in the question if the mental can cause something via its physical realizer, but rather if the mental can cause something *via* being mental. Causing something qua physical seems to always bring us back to the same spot: the reasons are epiphenomenal because they cause something only in virtue of the causal powers of its neural realizers. And as we will see further on, this is a serious issue for Davidson's approach.

Problems for Davidson's causal theory of action.

The problem possibly most relevant to the topic of this dissertation, aside from accounting how the mental could have an impact on the physical, was discussed by LePore and McLaughlin (1985, p.6). As a counterexample to Davidson's argument, they offer a scenario where, an individual wants to drink something fattening, and he believes that this can be achieved by drinking milk. He also remembers that his doctor, during the last visit, told him to drink at least a glass of milk a day to avoid calcium shortages. He decides to drink milk, and this action is caused by the two former events: him wanting to drink something fattening and his belief that milk would do the trick, but a primary reason for which he drinks the milk is the fact that the doctor told him to. The agent in question generally avoids fatty foods, trying to lose some weight, and therefore would have avoided drinking milk, if the doctor wouldn't have suggested it. This leads to an issue since in this case:

[…]the agent has a primary reason R for performing an action of kind A, intentionally performs an action of kind A, and in which R is a cause of his intentionally performing the action. But R is not a reason for which he performs A. (LePore and McLaughlin, 1985, p.6)

In the context of free will, one of the problems brought up frequently is the issue of implicit biases, or unconscious influences that could have an impact on the decision we are making. Although not identical to the argument brought by LePore and McLaughlin, the problem of what happens when an agent has reasons for action, and those reasons are recognized by him as causal, but in fact these are not the reasons but some form of an implicit bias that ultimately caused our behavior, are shared.

Davidson himself considers several different problems that can be raised in connection to his position, as: (1) the primary reasons including attitudes and beliefs, are states, or dispositions, but not events, and non-events cannot be causes; (2) a cause must be logically distinct from the effect, and since this is not the case here, reasons cannot be causes; (3) the same reasons cannot be sure to produce the same action (against what Hume argues in connection to causality), since there are no laws connecting the two of them; (4) the knowledge we have about our own reasons is not compatible with reasons being causes, no induction or observation is involved and therefore they cannot be connected in an ordinary causal relationship (which is usually established through one of those manners). In the first case, Davidson argues that although

state or dispositions are not events,[…] the onslaught of a state or disposition is (Davidson, 1980a, p.12)

Since Davidson is an event causalists, on his account only events take part of the causal chain of events, and as we have seen in the introduction to this chapter, although there are other propositions, this one is most broadly accepted. The mental events such as beliefs can start at the moment we notice something we desire. Such events, although not necessarily measurable or observable, exist, and can be part of causal relationships. Moreover, often in physical events, usually not questioned in this context, states and dispositions are often used as causes (e.g. *[…]the bridge collapsed because of a structural defect; the plane crashed on takeoff because the air temperature was abnormally high; the plate broke because it had a crack* (Davidson, 1980a, p.12)).

The second problem questions the existence of two different entities in the causal chain, where if reason explains an action by re-describing it, it would suggest that we are dealing with only one entity under different descriptions, rather than two distinct events. Davidson disagrees that this is an actual problem for his position, since:

To describe an event in terms of its cause is not to confuse the event with its cause, nor does explanation by redescription exclude causal explanation. (Davidson, 1980a, p.14)

Therefore, although it is possible to re-describe a cause, and describe an effect in terms of its cause, it does not entail that the two are the same.

The third problem is, in my opinion, linked to the position Davidson develops to explain how mental can be causal: anomalous monism, and to his approach to causality in general. Since Hume, it has been accepted that cause and effect are usually connected via a strict law, and while what we observe are just instances of co-occurrence, real causal relationships can be discovered because of the laws. Since there are no such laws connecting the mental and the physical, those relations cannot be considered causal. This issue, although not identical, is also closely related to the main issue discussed here, namely how can mental be causal in the physical world, since the physical is causally closed and the influence of mental does not seem to be governed by the same principles as the physical. This, along with Davidson's account of how this could be possible, will be discussed with more detail in the following chapters of this dissertation, but to deal with this specific instance of the problem, Davidson argues that:

The laws whose existence is required if reasons are causes of actions do not, we may be sure, deal in the concepts in which rationalizations must deal. If the causes of a class of events (actions) fall in a certain class (reasons) and there is a law to back each singular causal statement, it does not follow that there is any law connecting events classified as reasons with events classified as actions- the classification may even be neurological, chemical, or physical. (Davidson, 1980a, p.17)

The solution, rooted in Davidson's token monism, distinguishes the token individual causal statements, ones that can be backed by laws, with type classes which are not reducible, nor connected to laws that would govern them. Again the specificities of this approach are better described in the part of anomalous monism.

As for the last problem, Davidson argues simply that, although those classical methods of discovering causal links are effective, there is no reason to think that they are the only methods applicable.

Another major issue for Davidson's account, and possibly for all causalists, is the problem described by Sartorio (2009) as the problem of omissions. The issue stems from the types of actions conducted unintentionally, and it has been a notorious problem in philosophy of action. It is closely related to the problem of absences, where events are caused by the absence of something, for instance, when I arrive late to work, because the bus did not arrive, already discussed in the previous chapter, since it is an issue for all production accounts. Omissions are problematic for causalists, because they claim that only events are causal, and omissions are clearly not events, therefore Davidson, and others, cannot account for intentional omissions. The usual manner in which causalists reply to the problem is by stating that not only events, but possibly also other things can be causes and effects. The causation involving facts, for instance, is the most primitive kind of causal talk and event causation is in fact justified through its connection to facts. The agent then omits to do something, because he formed an intention that has caused the non-movement of the body.

Some, e.g. Davidson himself, endorse a view where the event causation is the basic kind, and other types involving other relata of the causal relationship, are made true by the causal talk that nonetheless involves events. In this case there is a mental cause that caused the omission, and if such mental event was absent, another mental event would take its place, triggering an action (possibly the one that I have failed to do). In the specific case of free will, Carolina Sartorio (2016) argues that free action is caused by both the events that produced the effect, as well as omissions.

Yet another possibility is to accept that there is more than one concept of causation, and omissions are causes of certain effects only under one of the possible concepts. As we will see in the next chapter, thinking of causality in a number of way is not uncommon. Some consider causal connection as productive, while others think of causes in terms of counterfactuals or difference making.

Where does this bring us?

Above I have discussed two more classical approaches to the action explanation, and before we move forward, to the contemporary attempts, a small conclusion might be in order.

The issues discussed above can be divided into a few categories:

What is action?

The question of the nature of an action, although seemingly very basic, seems too, often, be taken to be unambiguous and self-explanatory. This is particularly true when we look at the discussions in the experimental realm, where the understanding of action as an overt movement does not seem to be questioned.

In the paragraphs above, I have discussed views where the simple overt movement does not constitute an action. Although those seem to be in minority, they bring an important question to the foreground- what constitutes an action. For Anscombe the intentional action will have then at least two components, the intention and the movement, none of which will count as an effect of the other, therefore the intention potentially can be viewed as a cause of a movement, but independently the movement does not count as intentional action. The latter does seem to have mental causes though. It is not to say however, that movements of our bodies do not have causes. As Alvarez (2013) explains it, said movements are most likely caused by some, possibly very complex, neurophysiological processes. But, such things themselves are not actions, but rather consequences of actions, things that happen when and because we act. If we transfer this onto the discussion on the problem of free will, we can see that when action is considered in Anscombian way, the experiments are no longer valid as they are performed on the consequences of our actions, rather than actions themselves. Therefore the conclusions: the action was caused by a physical, unconscious process rather than the mental conscious decision no longer holds. Moreover, the fact that the movement had a neurophysiological cause is perfectly in line with the position and possesses no threat, if we still hold that free will is linked to an action being caused in a specific way.

For Davidson on the other hand, as is generally accepted, the overt movement, as is used in the free will experiments, is a type of an action and as such is caused by either a mental or a physical event. To be an action is, for Davison, to be produced in the right way, and more specifically by reasons, therefore the causal source matters already in identification if an event is an action, not only if it is an intentional action. This would differentiate actions from events that are not actions, and possibly differentiate between different types of actions (e.g. intentional and not intentional). The latter is problematic, as the same action, like for instance switching the lights on, can have different descriptions, like illuminating the room. For Davidson, in such cases, there is just one action that can be given a variety of different descriptions. This approach is deemed to be a coarse-grained one, in opposite to the fine-grained, where in cases like the one above, those are different actions. The former problem, how actions differ from non-actions, seems to again go back to the way they are caused, with actions being caused by a pair of belief an desire that constitutes a reason.

It seems that, for it to be relevant for the problem of free will, the actions under consideration cannot be neither automatic nor unintentional, as, intuitively, we would not call those free. For Davidson, it is logically impossible to perform an intentional action without appropriate reason, suggesting again, that reasons are somehow causal in the action generation.

What is causality, and is causal explanation quoting causes of the action?

Again the two positions discussed above lead to contradicting results, for Anscombe the reasons for action, referred to in causal explanations of an action, are not in fact proper causes. Anscombe moreover thinks of causation as a very general, umbrella term, that in itself is not particularly precise. She points out that there are much more detailed terms that can be used, each falling under the causality umbrella, that are nonetheless much more informative.

The core ideas propagated by Davidson, had humean understanding of causality at their core, where explanations of actions that cite reasons for which the individual acted are causal because they depend on the belief-desire pair that in fact causes the action. On Davidson's account (and we will see this is also true for Mele's account of intention) only the motivating reasons, and not the justifying ones, can be seen as the actual causes. The action was then an event identical to a bodily movement, caused in a specific way, which relied on a combination of a belief and pro-attitude.

What do we talk about when we talk about intentions; contemporary discussions.

From the discussions presented in the preceding parts of this chapter, two major options emerge to the question of what kind of causes are taken to cause action, and more specifically free action: reason and intention. In the contemporary discussions it is the latter that takes precedence, however it is often discussed using the model developed by Davidson, rather than the one of Anscombe, therefore the Causal Theory of Action. The question of what constitutes an action remains unresolved, with some arguing that more than just a movement, a bigger piece of the causal sequence, should be defined as an action (e.g. Pacherie, 2008)

The notion of intention appears frequently in philosophical, psychological and neuroscientific discussions, and again, much like other notions that are also used commonly, all three seem to use different understandings of it, and they differ in scope and depth, and it is difficult to provide a 'one fits all' solution here. The connections between intentions, action generation and free will are clearer than those of reasons, also more often discussed.

However dominant in the discussions, intention itself is not taken to be causal on its own, but rather leads to a formation of the conscious decision, which then serves as a basis for the execution of the intention, or it has some motivating component, that serves as the 'trigger' in the action execution. The relationship between intention, decision and action, not always clear, can lead to clarification of one of the issues in action execution, which Searle (2004, 2010) calls 'gaps'. Such gaps include: (1) the gap of decision making, since if connected to deliberation decision making is extended in time and ends with conscious decision; (2) the gap of the onset of the action, because having made the decision by itself is not sufficient for the onset of an action; and (3) gap of continuation of the action, since most of the actions are somehow extended in time, and all of the parts of the action form a continuous conscious voluntary action. Existence of those, suggest an existence of a continuous process, which could be linked to, with some parts identical with, the process of intention forming, and the intention could possibly be responsible for filling the gap between the decision and the actual execution of the action (mostly problematic in cases where the time lapse between the decision and the action is large).

In the free will debate, out of many different aspects of intention, its conscious nature is often taken to be central, and very problematic. It is often assumed that for an action to be considered free, the intentions and decisions that play a role in causing it, should be conscious. Even here, different concerns can be identified, and Mele (2009) divides the issue into two separate questions:

[…](Q1) Must I be aware (or conscious) of my intention to A for it or its physical correlate to play a significant role in producing an intentional A-ing? […]

(Q2) Must I be aware (or conscious) that I intend to A for my intention to A or its physical correlate to play a significant role in producing an intentional A-ing? (Mele, 2009, p.23)

In other words, for me to go to the kitchen to drink, must I be consciously intending to go to the kitchen, and do I also have to be conscious of the conscious intention for me to go to the kitchen. The two questions also emphasize the existence of the connection between the mental and the physical: the action is caused, or produced, by either an intention to A, or by the physical correlate of the intention to A.

Another problem, that appears in connection to intention and both action generation and free will, is its specific role in the production of the voluntary action. Again, often Davidson's position is used to explain it, where:

[…] an intentional action (an action done for a reason) is an event (a movement of the agent's body) caused by a mental event or state (a reason) and that, therefore, explanations of intentional action in terms of the agent's reasons are causal explanations (Alvarez, 2013, p.141).

As we have seen, Davidson initially rejected the causality of intentions, in favour of reasons, but later changed his mind, not being able to reduce intention in action to intentional action. We can therefore expect that an intentional action, which here would be interpreted as both voluntary and free, has some mental event as its cause, at least for causalists. For Davidson the appropriate mental event to cause such an action is a reason, but for others causalists, such as already quoted Mele, the same role seems to be played by the intention.

In some studies, intentions are investigated, but not specifically defined. We have seen this problem also in the free will debate, where in a majority of experimental approaches neither free will, not conscious decision, are explicitly defined. Some studies however take the proper definition of intention to be a crucial part of the studies, e.g. Haggard and Pacherie, define intention as:

[…] a mental state, which may be associated with particular brain states (Pacherie and Haggard, 2010, p.70)

Again, the mental nature of intention is emphasized, along with the connection between the mental and the physical, which suggests a dualistic approach to the problem would not be acceptable. The nature of this connection is however unclear, and a variety of problems can be identified, including if one intention has one, specific mental state associated to it.

Along the definition of the intention itself, the authors focus also on where our knowledge of intention comes from. While most intuitive would be to focus on agent's reports about intention, those are very problematic, as there is no way to check those subjective feelings of wanting to do something. Moreover, as we have seen before, there are deviant situations, where having an intention, although recognized by the agent, was not what ultimately caused an action. Haggard and Pacherie focus on the physical data instead, and they argue that:

[…] the most informative data [about intention] *comes from reports of direct cortical stimulation prior to neurosurgery for epilepsy* (Pacherie and Haggard, 2010, p.71),

and that such stimulation, of various regions of the brain, offers reasons to distinguish between initiation of action and conscious, immediate intentions, which Mele would call distal and proximal intentions respectively. The data seem to further support the hypothesis of a connection between the mental and the physical, possibly providing further reasons to reject dualistic solutions. However, also in those experiments, the nature of the connection between the physical and the mental realm still remains unclear, most likely involving some form of realization. The authors argue that a direct stimulation of presupplementary motor area is accompanied by an experience of immediate attention, while stimulating the deeper cingulate motor area

[…] produces a strong motivation to perform a specific action, and can trigger action initiation, but without any particular specific conscious experience prior to action (Pacherie and Haggard, 2010, p.72).

The intentions have, moreover, two distinctive features: (1) their connection to consciousness, and (2) their connection to the action, which could be its causal role played in action generation.

So what intentions are? Some, like Anscombe, emphasize the, epistemic, connection between the intention and self-knowledge, others, like Bratman (1987), see intentions as a way to coordinate actions, others, like Pacherie (2015), emphasise their role in controlling the actions currently underway. In all these accounts the role of intention is investigated, the what are they for question, rather than what they are. For the latter most agree they are, somehow, connected to the neural activations, however the actual relationship between them is often set aside. Pacherie (2015) states that the intention is

[…] a mental state that represents a goal (and means to that goal) and contributed through the guidance and control of behaviour to the realization of what it represents. (Pacherie, 2015, p.1)

Here again, the mental nature of intention is emphasized, but the focus is rather on what do they do, and if they are supposed to do something, i.e. have a causal role of some kind. For this dissertation however, the question what exactly it is that the intentions do is less relevant in comparison to the how are they causal.

The manner of causality, or how the intention is causally related to the action, can also be distinctive for two specific reasons: (1) a causal reason, and (2) a content reason (Pacherie and Haggar, 2010). The former treats intention, a mental state, as causing an action, potentially a physical state[9], and as such, as we will see with more detail in the following chapters, is susceptible to a number of issues, most pressingly mental causation. Moreover it can be questioned if the fact that intentions are useful in explaining the action is reason enough to constitute an evidence of either their causality or their ontological status as mental states. The latter, a content reason, treats the content of the intention as being connected to, some, details of the overt physical movement[10]. Therefore each action should, in principle, have a different content of intention, but this approach is potentially vulnerable

[9] It is a simple hand movement in the example discussed by Pacherie and Haggard, however, as we have seen before, on Pacherie's account this would not, in principle, constitute an action, since the latter is a combination of an overt movement and an intention.
[10] Again, in this specific approach the movement is treated as an action.

to some reductive problems, namely how detailed do those intentions have to be, and how different do they have to get: is grabbing a can in one way, with a specific velocity and grabbing power, connected to a different content intention than grabbing the same can with a different velocity? Pacherie and Haggard argue that

The content of intention should be discriminative, in the sense that it should predict specific details of action. The content argument emphasises the continuity between decision and intention: when someone decides to do A rather than B, they may develop an intention whose specific content will relate to A or to B. (Pacherie and Haggard, 2010, p.3)

Again, the issue how detailed the intention should be remains, and it seems that the argument is only useful in cases when two drastically different actions are chosen. Similar intuitions on intentions were also explored by Soon and colleagues (Soon et al., 2008) as well as Haggard and Eimer's (1999). In the former study, already discussed in the previous chapter of this dissertation, the actions the participants were asked to performed, involved choosing between left and right, and the working assumption was that depending which side will be chosen, the intention will also be lateralized.

The fullest and probably most exhaustive account of intention is offered by Mele, who links it explicitly to both the problem of free will, as well as to the conclusions of Libet's experiments, and according to him the results obtained by the latter, the readiness potential is not a marker of action initiation but rather could be a neural build-up that accompanies forming of an intention. Such approach would eliminate the conclusion that action is initiated unconsciously, and ultimately falsify the, often drawn, conclusions based on the experimental results, pointing to the impossibility of free will. Mele divides intentions into different categories, depending on the role they might play in the action generation and their temporal distance between the two.

Occurent and standing intentions.

First division, onto: (1) standing, and (2) occurent intentions, is based on the relationship between intention and action, and the existence of so-called executive attitudes towards the action. The two types should not be taken as, disconnected, opposites, but rather as two different sides of a spectrum, with different grades between them.

Standing intention, as defined by Mele, is a disposition of a certain kind to have corresponding *occurent* intentions (Mele, 2007, 2009). The standing intention can be attributed to us, without questioning if at the time of the attribution we were conscious of possessing that intention or not. The standing intention can be illustrated through a plan to go on holidays. Let's say that we have an intention to go on holiday in August, and tell somebody, a friend, about said plan. The friend then, in another conversation with his wife, tells her about our plan, assuming that we still have such intention, regardless of the current state we are in (we could be sleeping, or simply not thinking about holidays, focused on our work). Therefore a standing intention is our ability to have an occurent intention, which is defined as:

[…] executive attitudes toward plans (Mele, 2007, p.741; also discussed in Mele, 1992 and Mele, 2009)

he standing intention is then a disposition of a specific kind, where an agent has, or is in, an executive attitude towards a specific plan. It is our ability to plan, and decide to execute a specific action, encompassing both immediate plans as well as ones regarding an action further away on the timeline. The spectrum of states, between the two intentions, could potentially account for the gaps identified by Searle, especially one between forming of an intention and the action. The gap could be a difference between being in an active, occurent intention (in the moment of forming, and just before the action is executed), and the somewhat passive, standing intention (where the intention is still present just not attended to). Which could possibly result in a scenario illustrated below:

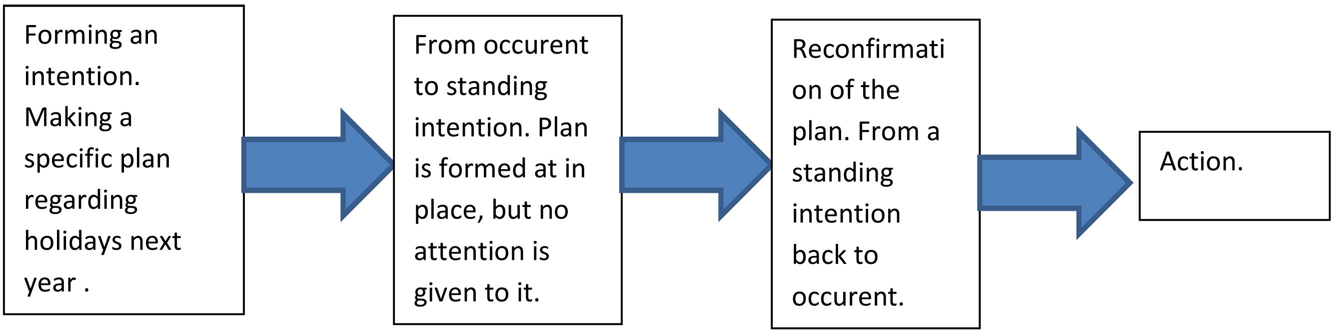


Figure 11. Action generation: the interaction between the standing and occurent intentions.

The illustration above is just a rudimentary proposition of how we could account for Searle's gaps in action generation using Mele's approach, but it shows that an action could be caused by an intention that, as time passes, changes its characteristics, based on the amount of attention that the agent is giving it.

For the intention to be occurent, it has to fulfil one of the two possible conditions, firstly it can be:

[…]suitably at work at that time in producing relevant intentional action or in producing items appropriate for producing relevant intentional actions (Mele, 2007,p.740).

Here the occurent intention, as the executive attitude, takes part in the causal chain of events that eventually lead to the action. Again, it has to cause the action in the right way (has to be suitably at work). Interestingly, the occurent intention does not have to itself be the immediate cause of the action, it can also play a causal role in producing other events (or items) that will then lead to the intentional action (I assume Mele means here that those events will then become the causes of actions).

Secondly it can be:

[…] not wholly constituted by a disposition to have occurent intentions to A (Mele, 2007, p.740)

One occurent intention can fulfil both conditions at the same time, although it is not necessary. However, to be classified as occurent, the intention has to at least fulfil one of the conditions.

Similarly to beliefs and desires, also intentions have representational content, which, in the case of occurent intentions, is an action plan. The representational content of the attitude and the psychological orientation are also distinguished from one other, since:

Orientations include (but are not limited to) believing, desiring and intending. On my view, the executive dimension of occurent intentions is intrinsic to the attitudinal orientation intending. (Mele, 2009, p.6)

Plans for the execution of the intention themselves can be divided into simpler and more complex ones, or, as Mele defines it, having one or multiple 'nodes', with the higher amount of nodes signifying a higher level of complexity of the plan. To have an intention, or to have an intending attitude towards a specific plan means that an agent has set his mind on the execution of an intention, and is going to go through with the plan to put if colloquially. This broad approach to how intentions, and action plans work, allows to accommodate very loose plans (where an agent has an approximate idea of how he is going to proceeded), as well as the agent changing his mind. The plans, on Mele's account, range

[…] from simple representations of prospective "basic" actions to complex strategies for achieving remote goals-constitute the representational content of occurent intentions. In the limiting cause, the plan component of an intention has a single "node". It is, for example, a prospective representation I have of raising my right index finger[…] (Mele, 2009, p.5)

The more nodes, the more complex the plan is, and those can range from very siple to very complex. Plans and occurent intentions are much more action oriented than desires, since having the latter to go on holidays seems to leave the action execution open, while the former seems to include the executive attitude, which the latter lacks. In order for an intention, and specifically occurent intention, to fail, seems to require an appearance of some, unforeseen, circumstances or external factors that disrupt the action execution. In case of desire on the other hand, not executing of the plan seems to require nothing more than changing one's mind, as desire does not seem to share the executive attitude with intentions. Here Mele's position very much resembles Davidson's approach, who also deems desires to be insufficient for the action execution, and reasons, ultimately causal in action generation, require both a desire as well as an action oriented component.

In the generation of action, as relevant to the discussion on free will, an intention is connected to a decision, and in the neuroscientific experiments it is the latter that is causal. Mele defines decision as

[…] to decide to A is to perform an action of forming an intention to A (Mele, 2003, chapter 9[11]),

This is significant for a number of reasons. Firstly it further emphasizes the tight connection between intention an decision, but secondly, it defines a decision as a type of action itself. This could possibly be interpreted in two ways: the intention is a cause of the decision, or the decision is identical with the forming of an intention. The latter would emphasize again the executive nature of intentions, while the former would focus on their causal nature. However when we analyse how Mele uses the argument to disarm the usual conclusions of Libet's experiments, the former interpretation seems to be the correct one, where he argues that the RP is a preparatory build-up of neuronal potential connected to the process of intention forming. The moment identified by the agent as the conscious decision, is in fact the intention reaching the final stages of forming, or being formed.

One way for the individual to change his mind, after forming of the intention/making the decision, is to acquire a belief that he is not going to do A (Mele, 2009). Then two intentions, to A and not to A, seem to enter in a competition, which is ultimately won by the intention not to A. It is unclear which factors contribute to one of the processes being chosen, but some light on this can be shed by the dual process theory, which will be, briefly, explained further in this chapter. Also here desires seem to differ, since we do not usually form two competing desires. Intentions, specifically the occurent ones, are also not to be understood as a combination of desire and belief, although there are connections between all of them, making a clear difference between Mele's and Davidson's position.

Occurent intentions are not necessarily rigidly linked to the act of deciding, since it is possible to have an intention and act upon it, while no decision has been in fact formed. This is usually the case with acts that are conducted out of habit (like turning a key and unlocking the door, you have an intention to do it, and in normal conditions there is no need to decide

[11] For this reference the version of the book from Oxford Scholarship Online was used, therefore the exact page is missing.

to unlock it). Habitual actions are an issue for theories of voluntary action, as they appear intentional, yet they are not caused by a conscious decision in the right way, e.g. when we get home from the office in the evening, we usually open the door, but seems very rarely make a conscious decision to do so.

Proximal and distal intentions.

Another way to distinguish between various sorts of intentions is to divide them based on their proximity to the action itself. More long term plans, and intentions that are about actions that will happen in a more distant future (e.g. next week, next month, next year etc.) are distal, while the once concerning immediate action are proximal. The latter are also responsible for the continuity of action (e.g. I will continue to write this dissertation in the immediate future). There are also intentions that are a mix between both the proximal and the distal intention, when an action that we attempt to execute will start immediately but will take a while to finish. The division onto the two is not a strict one since it is hard to find a precise point in which an intention stops being proximal and becomes distal, as an action that we will conduct later might mean in six minutes as well as in six weeks. The fluidity of intentions, as far as their temporal relationship with the action goes, brings back the issue discussed by Searle, regarding the gaps between forming an intention and acting. The relationship between the distal intention and the action seems particularly relevant in this light, especially if we consider distal intentions to possess the same characteristics as the occurent intention, namely its executive nature.

Three tiered dynamic model of intentions.

One of the more significant models, connecting both the empirical and theoretical discussions on free will, and those on action initiation, is the three-tiered model developed by Pacherie (2008). She starts from the many different components of action, such as awareness of a goal or intention to act, and develops a way to connect those, and their possible sources, and through that explain how action might be caused. The reasons for this new division are twofold: (1) it seems that the intentions come after each other in a sequence, where the moment the second one begins, the first one ends, and (2) the control over the action is given primarily, and almost exclusively to the second indention, closer to the action itself (on Mele's model, this would be the proximal intention).

Pacherie divides the intitions onto three catgories: (1) D-intentions, or distal intentions, which seem connected to the deliberation and planning; (2) P-intentions, or proximal intentions, which are linked to the intention to act now, up to the completion of the action, and (3) M-intentions, or motor intentions, which neuroscientists call motor representations. Those intentions are responsible for the organization of motor sequence.

Those three work in sequence, however when all active, they can work in parallel, thus the forming of an M-intention, does not mean that either D-intention or P-intention became non-active. The sequence can be illustrated as follows:

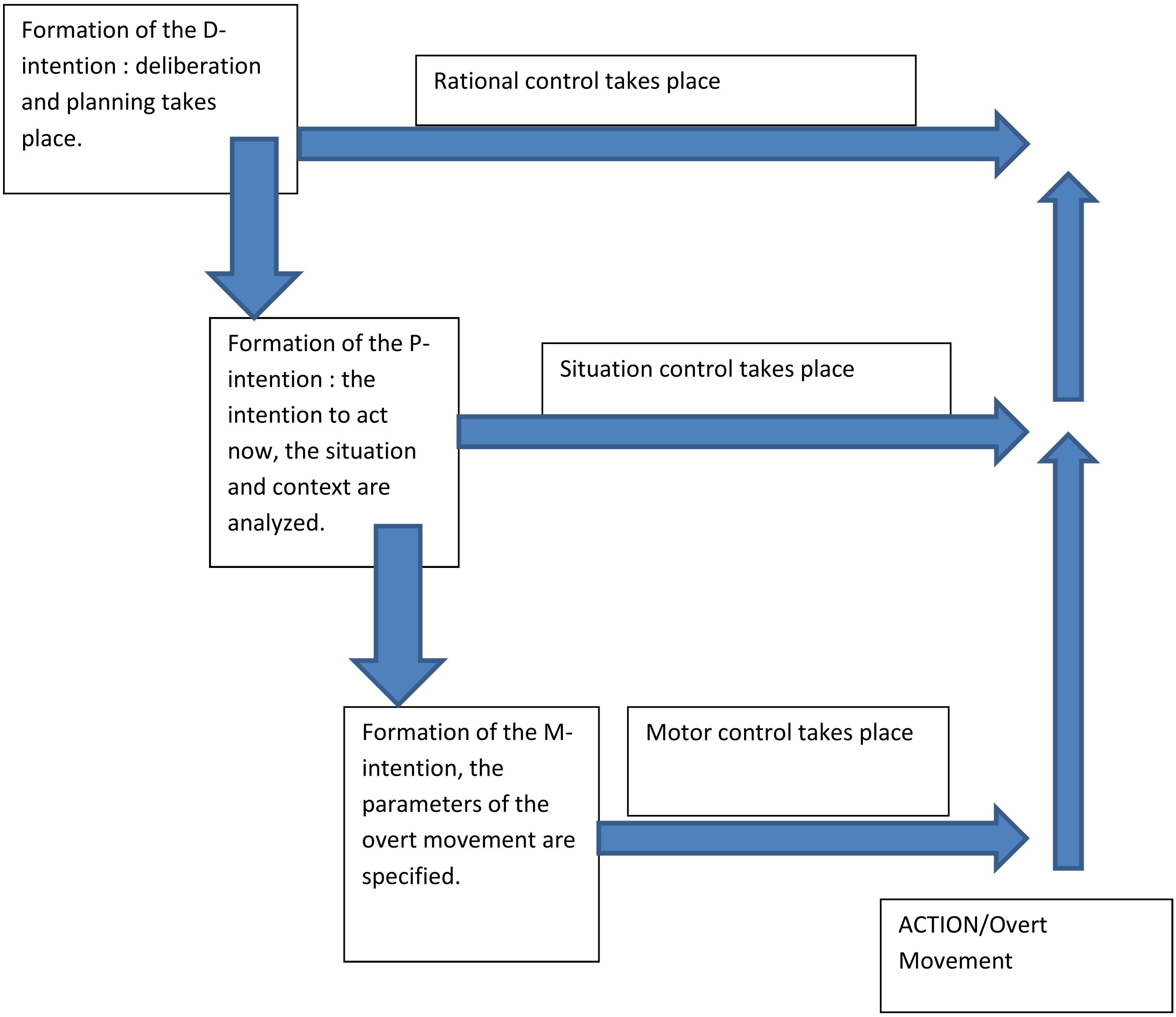


Figure 12. Pacherie's model of intentions, adapted from Pacherie (2008, p.189).

As we can see on the illustration above, Pacherie's model not only assumes that all the different intentions are at work at the same time, but also differentiates between the roles they play both in action execution as well as in controlling of both the action and the movement. The word action on the scheme above might be a bit misleading, since, as was mentioned before, Pacherie follows philosophers like, e.g. Searle (1983), who argued that action consists not only of the overt movement, but additionally an intention in action that caused it. This seems to follow opposite intuitions as those of Mele, who seems to consider simple overt movements to be an example of an action. In both cases however the intentions have two main, essential here, characteristics: (1) they are mental, and (2) they are, somehow, connected to the physical realm.

What is the connection between the intention and the physical realm?

As we have seen throughout this subchapter, some form of a causal role played by intention in connection to action is usually assumed. This is linked to several, discussed previously, issues, including the identification of action (as an overt movement, or broader), and the exact role that the intention can play (as a tool for explanation or rather genuinely causal). What is often omitted or discussed only briefly is the connection between intention and the physical realm, as mentioned several times throughout the chapter.

If we assume that intention is somehow, genuinely, causal, as is often done in the discussions on free will and intentional action, the connection between intention and the brain needs to be addressed. In most cases a dualistic approach is rejected, and some form of dependence is assumed, under which it is argued that

Whenever human beings perform an overt intentional action, at least one of the following plays a causal role in its production: some intention of theirs; the acquisition or persistence of some intention of theirs; the physical correlate of one or more of the preceding items. (Mele, 2009, p.11)

Here Mele clearly assumes that a (free) intentional action is produced either though an intention, intention acquisition or through the physical correlate of the intention, providing further reason to believe that: (1) intention has to be causal in the action production either by itself or through its physical realization or correlate, and (2) the two are somehow connected (since their causal influence on the action is exchangeable). Therefore we can argue that Mele is in favour of some form of physicalism, either reductive or nonreductive. The reductive version, which would assume that the mental can be somehow reduced to the physical, would however eliminate the duality of causes, the mental intention and its physical correlate, and only the latter would be in fact relevant for the action while the former would be reducible to it. Therefore it is much more likely that Mele is advocating some form of nonreductive version of physicalism, where for an action to be (free) intentional, it has to be caused by either one of the connected causes: the mental intention or the physical brain process.

The other models of intentional action, by Pacherie, Davidson and also Anscombe, similarly advocate some form of the position, where intention, while being mental, is somehow connected to the physical realm. They therefore seem to be far from concluding that two competing substances exist, and cause action. Again, while this is not the case, most of the positions are compatible with dualism, and except from Davidson[12], none of the authors analysed here provides a complete account of how the two are connected (e.g. what kind of relationship links them), nor how would the mental be causal in the production of a physical action[13].

Intention and neuroscience.

Aside from the discussions in philosophy, intention has been also increasingly more interesting to neuroscientists. Significantly, some (e.g. Haggard, in Pacherie and Haggard, 2010) discuss Libet's experiments as involving some work on intention, although they emphasize that Libet himself has been avoiding the word itself. This brings us again to the conceptual issues with exact definitions of what is Libet investigating: is it an urge, a decision or something different entirely. Moreover, here the issue of the time lapse between the intention and the action is particularly problematic, since Libet's approach only allows to monitor a very short window, in which the intention would be causal, and, as pointed by Gallagher (2006), this might be too short of a time for what we usually expect from free action, as such might require temporally extended processes of deliberation. In the theoretical domain this has been, at least partially, remedied by Pacherie's model (2008), described above.

[12] Davidson's position on this: anomalous monism, will be discussed in detail in the next part of this dissertation.

[13] Or action that is partially physical. Again the interpretation here is contingent on what we consider an action to be, just an overt movement or something more. Since one of the aims of this dissertation is to provide an interdisciplinary approach, and neuroscientists seem to agree that action is the overt movement, this is also the stand I will, mostly, adapt here.

Equivalently to Mele's position, the neuroscientists usually recognize some form of connection between the decision and the intention, where the latter leads to the former. If we follow this line of thought, parallels can be seen between Pacherie's model and Brass and Haggard's (2008) model of intentional action, where three components of the decision, ultimately leading to an action, are singled out: (1) the 'what', (2) the 'when', and (3) the 'whether' decision.

Brass and Haggard define intentional action as having a purpose and coming from, or initiated by an agent rather than something external (therefore is not stimulus-driven). This, brief, definition has several problems. Firstly the purposefulness of the action seems to be a complicated notion, and can mean anything from being goal-oriented (as I will treat it here), to having a reason for conducting it (connecting it to the work of Davidson). Secondly, being initiated by an agent could possibly suggest that an agent is a separate substance, capable of causing something. Since this does not seem to be the aim of neuroscience, I will treat it as an event that the agent recognizes the authorship of, still problematic, but superficially will serve its purpose. Further however Brass and Haggard argue that:

[…] the nervous system must make decisions, or generate additional information to produce intentional behaviours, which is not required for stimulus-driven action. (Brass and Haggard, 2008, p.319)

Although there can be multiple issues with the interpretation of this definition, I will take it to mean that the intentions are treated as, somehow, connected to the physical nervous system, although I think that it could be also taken as a reductive approach.

The generation of information as leading to the production of intentional action has, according to Brass and Haggard, three key components: (1) the 'what' component, responsible for deciding which action to execute[14], (2) the 'when' component is associated with the timing of the action execution, and was investigated by Libet (et al. 1982, 1985), and (3) the 'whether' component, which is linked to determination if any action should in fact be executed, or no action at all. The last one could, potentially, be connected to Libet's veto, where the decision is made if the action should continue, or it should be abandoned.

[14] It is not clear however how that decision comes to be, nor if it is connected to actual possibilities truly existing. It has to be however signalized here that the discussion on the what component could be liked to the problem of alternative possibilities. Moreover it seems that Frankfurt's solution would not work here.

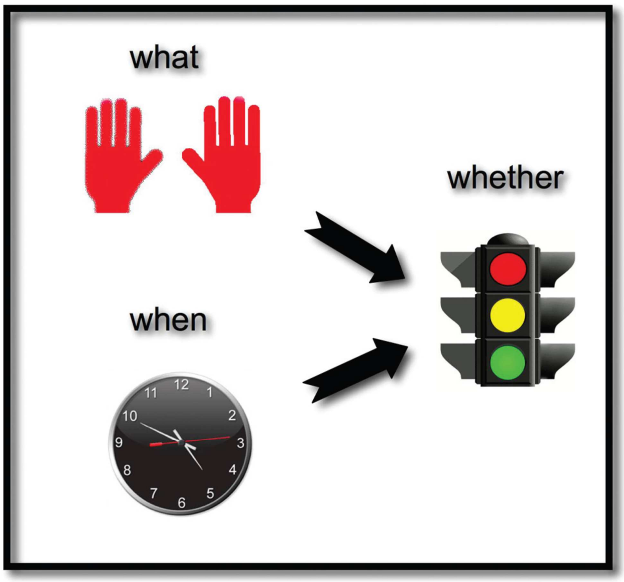


Figure 13. Schematic presentation of the what-when-whether model from Brass and Haggard, 2008, p.320

The tests on the first, ‘which’, component are done primarily by comparison between stimulus-driven, externally triggered action, and actions where participants can freely choose between alternatives (e.g. Muller et al, 2007). The studies have found that:

When comparing brain activity in a condition that requires the free selection of action with a condition where the responses are externally triggered, activity in different parts of the fronto-median wall has been reported. The most consistent fronto-median brain activity was located in the rostral cingulate zone (RCZ) and the pre-SMA. (Brass and Haggard, 2008, p.320).

This finding would potentially allow for a method of differentiation between the areas that are the physical realizers of those specific types of intentions, and those which are not. However, there are obstacles to such interpretations, as the brain activation might also be connected to conflict resulution while chooosing between alrenatives, rather than by the intention to choose one action. Also methodological obstacles can be easily identified, as experimental setting brings about set of rules that participants need to follow, that might, ultimately impede the participant’s choice mechanisms.

The second component is related to the problem of timing in causal mechanisms. Libet interpreted the timing of the physical processes, as prior to the conscious decision identified by the participant, to mean that the action is not consciously caused, but rather is made by some unconscious physical mechanimsm. Brass and Haggard (again arguing against dualism) argue that intentional action in its timing aspect, can be characterized by

[…] looser connection to external events compared to stimulus-drivern action, by their relations to complex contextual patterns rather than identifiable single stimuli, and by their mediation via memory traces rather than immediate stimulation. (Brass and Haggard, 2008, p. 321)

This approach steps away from the Libety-style paradigm, where the decision seems to appear from nowhere, not connected to the physical process althoghether, a position that would be just as difficult to accept for philosophers. However, they still hold that the timing of the onset of brain process can nontheless can provide insight on free will. However some (e.g. Schlosser, 2014) have argued that establishing the temporal order is not sufficient to claim any conclusions regarding free will.

The 'whether' component of intentional action is probably least investigated, as it requires some form of inhibition processes. As was discussed multiple times in this dissertation, the neuroscientific methodology is constructed in such a way, that makes an overt movement an important feature of the measurements. Therefore research on withholding an action is particularly challenging. Brass and Haggard (2007) have conducted an experiment using Libet's paradigm, where they were asked to, if they wish, stop the action underway, just before they would press a button. Again, as in the previous components, the brain activation found (doors-fronto-median cortex and anterior insula), supported the differentiation made by the authors onto the three components of intentional action, which differentiates them from stimulus-driven actions.

The three components are related to, although do not seamlessly reflect, Pacherie's model of intentional action. My underlying assumption is that, following Mele, also for Haggard and Pacherie the decision is a formation of an intention to act, therefore the discussion on intention can be, somehow, mapped onto the discussion on decisions. Haggard's division seems to reflect the types of intentions more closely connected to a specific action than the D-intention, which seems to be more abstract. None of the decisions discussed by Brass and Haggard seems to reflect the rational part of deliberation, but it has to be noted, they were not aiming at explaining the long-term, forward planning, that is typical for the D-intention. Their approach can be seen as potentially suggesting that the P-intention can be further divided, as all three of the decisions identified by Brass and Haggard seem to be related to the latter. However similarities can be observed between the D-intention and the 'what' decision, as both seem to be connected to the process of deliberation. The latter seems to also possess a component typical for the P-intentions, namely the contextual component of the action, where the 'what' decision is made, after the state of the surroundings is updated and the possibilities are identified. As for the physical anchoring of the 'what' decision, the experiments are conducted chiefly on patients with brain damage.

Patients with damage to the ventromedial prefrontal cortex (VMPFC) are often described as having impaired ability for planning and decision-making despite retaining intact capacities for explicit reasoning.[….] the deficits of VMPFC patients result from [according to studies by e.g. Bechara et al., 1999] *a failure to link an implicit emotional response- a somatic marker- with an explicit representation of a situation[…]* (Pacherie and Haggard, 2010, p.13)

This would further emphasize the connection between the 'what' decision and the P-intention, as its physical basis connects it to the representation of the situation, which is the domain of the P-intention.

Similarly the 'when' and 'whether' could be linked to the P-intention, as they all require assessment of the situation that the individual is in, and both seem to be about implementation of the intention. 'When' would be the type of intention targeted by Libet, and it clearly is connected to the action, suggesting some connection with the motor intention as well. The 'whether' on the other hand seems strictly linked to the situational assessment, whether the action should be executed relies on the individual's goal, and the possibility of meeting it in a given situation, through a given action.

It is also possible to think of an intention as being more specific, but only one, and define it as an event or a process which

[…] specifies the goal of a movement and the type of movement […] (Andersen and Buneo, 2002, p.191)

Again, this binds intention directly with an action understood as an overt movement, however, on Andersen and Buneo's account, it is possible to have intentions without acting upon them, therefore without actually moving. This would potentially deprive intentions of their executive attitude, emphasized by Mele. On Andersen and Buneo's account intention and its neural correlate seem to encompass what Brass and Haggard divided into three intentions, without being very detailed, as they argue

[…] a neural correlate of intention does not necessarily contain information about the details of a movement, for instance the joint angles, torques, and muscle activation required to make a movement […]. Intentions are initially coded in visual coordinates in at least some of the cortical areas within the PPC [posterior parietal cortex]. This encoding is consistent with a more cognitive representation of intentions, specifying the goals of movements rather than the exact muscle activations required to execute the movement. (Andersen and Buneo, 2002, p.191)

Therefore the intention is not related to, or identical with, a specific, detailed plan of the movement, but a very general one. As we can notice, again the term neural correlate of an intention is used, but the exact relationship between it and the intention itself remains unclear.

The various neuroscientific approaches to the notion of intention seem to emphasize both its causality, as well as the connection between the intention and the physical realm, while retaining the mental nature of the former. This seems to fall in line with the approach to intention found in philosophy.

The location of the brain region responsible for, either, intention as one action directed process, or the various types of intention might be problematic, as Brass and Haggard argue intention

[…] is not generated by a single brain area but is a product of a recurrent fronto-parietal network (Brass and Haggard, 2008, p.323)

This however only further emphasizes the complexity of the phenomena in question, and puts into doubt the results obtained either using experiments with limited time (e.g. Libet), or ones that investigate only specific areas of the brain (e.g. Soon et al., 2008).

The experimental focus on the immediate attention, mostly motor ones, was, as we have seen previously, criticized by Gallagher (2006), who argues that free will cannot be seen as being part of only those, limited, decisions, but rather comes with more abstract, complex processes. Both Pacherie's account as well as Brass and Haggard's could be seen as a remedy to this issue, however it is still unclear how the distal intention would be an effective cause of an action.

More on the differences and similarities between intention and decision.

The difference between the two: the intention and the decision, is significant here because of its connection to the experimental research, where intentions, desires, beliefs, urges and decisions are often mixed up[15].

In the context of action generation Mele has provided with the fullest account of the relationship between the intention/decision pair and the different types of decisions, although as we have seen, ultimately deciding is forming an intention to perform an action.

The first distinction divides deciding onto practical and cognitive (Mele, 2007). The former is strictly linked to action, and both deciding to do something, e.g. go on holidays, as well as deciding not to perform an action, e.g. deciding to skip holidays this year, fall under this category. The cognitive deciding on the other hand is not connected to actions, and:

[…] *to* [cognitively] *decide that p is the case is simply to acquire a belief that p is the case of the basis of reflection* (Mele, 2007, p.3)

In connection between practical deciding and the formation of intention can be then defined as follows:

[…] to decide to A is simply to acquire an intention to A on the basis of practical reflection, and acquiring an intention- in this way or any other- is never an action. (Mele, 2007, p.198)

[15] As I discussed previously this is particularly true about Libet's research where the decision to move is mixed with urge to move and the feeling to move. As we have seen in the chapter on neuroscientific research on free will, this is one of the major problems of the approach.

The acquisition of an intention, as we can see here, is not itself an action. The practical reflection Mele is referring to

[…] *issues in evaluative beliefs about action: for example, the belief that it would be best to A, the belief that A-ing would be acceptable and the belief that A-ing is satisfactory means to a desired ends.* (Mele, 2007, p.198)

The reflection, which then forms the basis for the intention and decision is then, much like in Davidson's model, based on some form of, probably practical, belief, in which an action is deemed to be the best way to achieve a desirable outcome.

Mele, in discussion on decision, reconfirms its mental nature where he defines decision as: *[…] momentary mental action of intention formation.* (Mele, 2008, p.3)

This would then mean that all the potential causes of an action, discussed so far: intention, reason and decision, were identified as mental entities. All of those are also specifically put in position of causes of an action, and in connection to practical deciding Mele argues:

[…] actively settling on a course of action-may be understood as a mental act of executive assent to a first-person plan of action. (Mele, 2007, p. 210).

Deciding is then forming an intention, as was said before, and moreover the intention incorporated in itself a plan of action, of unknown level of specificity. This is valid both in cases where the decision and intention are followed by an overt movement, as is assumed in neuroscience, but also where the intention formed pertains to not acting. The latter, as the former, requires all three: intention not to act, decision and a plan.

The approach to deciding seems to bring Mele's position ever closer to the one of Davidson, where everything was caused by a belief/desire pair that constituted a reason for acting.

The missing agent.

The biggest problem in framing the problem of the causes of our actions in terms of mental events: intentions, reasons or decisions, is the problem of the absent agent. All the theories discussed above explicitly determine the causal processes as containing various, mental and physical, states or events, however none of those is identical with an agent, who presumably should be in charge of both the production and the control of the action or behaviour. In the experimental domain it was even explicitly stated that one of the conditions for an action to be classified as voluntary, the agent has to recognize that he was the causal source of the action, again suggesting that the agent is the cause of the behaviour, or at least triggers the causal process that results in the behaviour.

There are several ways to deal with the problem, depending on what we are willing to give up. Two versions, that try to keep up the essential components of the classical (davidsonian) approach, where an agent is either (1) treated as a passive participant and the actions happen to him, rather than are produced by him, or (2) the agent is eliminated which leads to:

*[…] the abandoning of anything resembling an account of agency understood as the capacity that subjects possess to give rise to and control an action. (*Aguilar, Buckareff, 2010, p.13)

Another way to deal with this problem is to develop another way of understanding how action is caused, where the agent is the ultimate cause of his own behavior: the agent causal theory of action (e.g. Alvarez and Hynman, 1998). This solution can also be found in the discussions specifically on free will, and it is considered an incompatibilist account of free will. The latter is based on the same underlying assumptions as the account in philosophy of action, where for an action to be free, it has to be caused by an agent and not by an event[16]. Here the agent has to not only be the cause of the action, but additionally the latter cannot be determined by the prior events combined with the laws of nature. The position was most fully developed by Timothy O'Connor (e.g. 2000, 2009), while Pereboom (e.g.2014) argued that although free will in fact requires that an agent is the cause of his own actions, the burden of proof is on those positions, as currently the evidence does not support this thesis.

Proponents of agent causation argue that in sentences like Sue broke the bike by hitting it with a stick is synonymous with one stating that Sue was the cause of the bike breaking. Therefore it was not an event that caused the bike breaking, like the stick hitting it,

[16] Most of the accounts of free will, also in the classical discussions, and in neuroscience, consider events to be causal.

but rather Sue, as an independent agent, temporally persisting. What about intentions though? Vihvelin (2013) describes the chain of events that leads to an action as involving intentions but ultimately the actual cause is identified elsewhere, in the agent, who is almost seen as the first unmoved mover. She states:

*Granted, Joe [*in the example used here, Sue] *caused the braking of the window by first causing some change in himself- his choice, decision, intention, or other mental act or event. But when we trace the causal chain of events backward from the broken window, to the rock's collision with the window, to Joe's throwing the rock, to Joe's choice, intention, or some other change in Joe, the chain of event causation comes to an end. Some event involving Joe- let's call it 'Joe's basic action'- has no event causes but is nevertheless caused. Joe- the temporally persisting person- caused this event. And no event (or collection of events or thing or persons) caused Joe to cause it*. (Vihvelin, 2013, pp. 57-58).

We can then explain the agent's behavior through the desire to break the widow, or the intention to do so, and we could possibly connect it to the davidsonian idea, where having both the desire, and the belief that a specific action would satisfy said desire was required, however the intention that ultimately led to the decision and the action was caused by the agent himself.

The main attraction of the position seems to be providing the type of control we seen to require when we talk about voluntary action, stemming from our beliefs and desires. Not only we have to be able to do otherwise, we also need to be able to control ourselves and our environment, which if we think in terms of chains of events, it is difficult to see what would possess such control. It also seems to be very intuitive, in the end we clearly see ourselves as acting agents, doing something in the world, based on some reasons or intentions, depending on which we find central. However, the account also possibly leads to a form of dualism, where the agent is treated as a separate substance, and as such seems to some extent incompatible with the neuroscientific approach to things.

Automatic vs. goal-directed processes.

In the previous sections of this chapter two types of actions were briefly mentioned, namely the stimulus driven and goal directed. Those two types of behavior, or action, form an opposition, and are usually described in terms of dual process theories, and they are usually described in connection to the mechanism of choice and therefore as goal-directed mechanisms and stimulus driven mechanisms of action selection. This two types of actions of behaviors are crucial for what we consider free movements to be, as it is almost unequivocally assumed that free action cannot be automatic, but rather slow, purposeful and based on deliberation. Moreover the stimulus driven actions are usually caused by some external causes, while it has been already often emphasized in the previous sections, the free and intentional actions originate from the agent, and are not a response to external stimulus.

The stimulus driven mechanism is often defined as:

[…] one that is mediated by an association between the representation of stimulus features and the representation of the response (an [S-R] link). (Moors, 2017, p.68)

Such mechanism can be understood as creating a pair stimulus-behavior, where, presumably every, time the stimulus appears the behavior follows.

Stimulus-driven behaviors are often defined as rigid and automatic, operating in conditions where there is no time for elaboration, and the stimulus is immediately followed by an overt response of the agent. As such it cannot be considered an example of free action, in fact under specific interpretation of stimulus, the actions chosen by Libet could be interpreted as automatic, and therefore not free.

A goal- directed mechanism, on the other hand is usually defined as one that is

[…] mediated by an assessment of the expected utility […] of action options. This assessment is based on the interaction between representations of the values and expectancies of the outcomes of an action options ([S:R-Ov] links). (Moors, 2017, p.68)

Therefore the goal-directed behavior seems to rely on making a choice between different action-options, and ultimately executing one with the highest expected utility[17]. This type of action seems to fall in line with what we usually consider free, and has the usual, often problematic, components, including the existence of different options to choose from, and the process of making a decision. Goal-directed mechanisms, unlike stimulus driven are said to be slow, but much more flexible, and therefore are assumed to operate under conditions where time is not a major factor (as they require deliberation). The identification of the goal directed processes seems to have, at least, some connection to the standard belief-desire theory, as discussed by Davidson. In such context, the action linked to the highest utility, then becomes the goal to act. This can be illustrated through the scheme below:

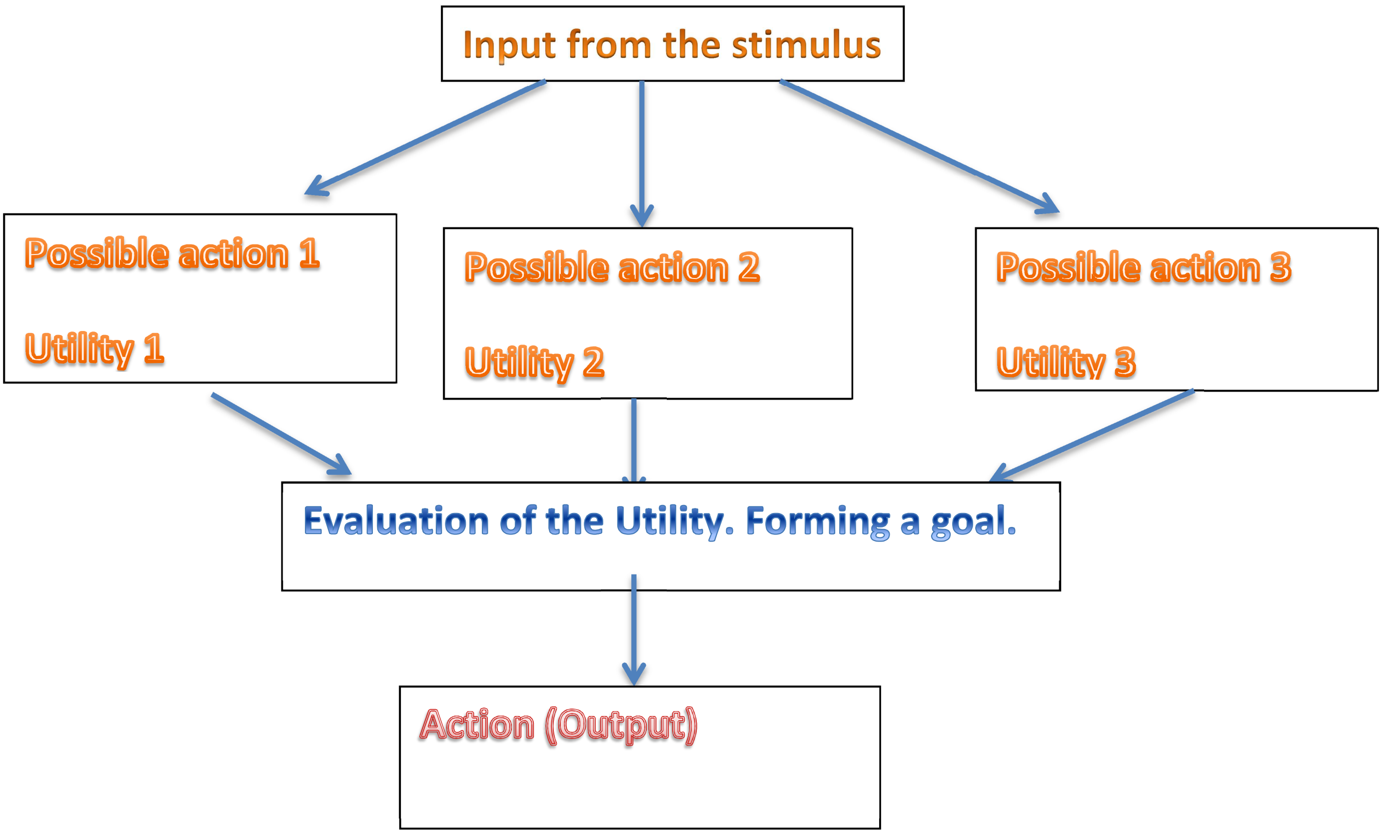


Figure 14. Schematic representation of a goal-oriented action initiation.

As we can see, a stimulus from the outside is still present, which may be interpreted as the context of an action, where we identify possible actions based on our sourrounding,

[17] For more on how the utility is assessed, e.g. Heyes and Dickinson, 1993; in a nutshell the assessment stems from the interaction between the representation of the values and the expected outcomes of the different actions, discussed also in Moors, 2017

therefore the notion of stimulus is not necessarily synonymous with the one used in the stimulus-driven action. We identify different action options, and then analyze their utility, and choose one that fulfills our initial goal best. In case of stimulus- driven action, the intermediate states are assumed to be skipped, and the input from the stimulus leads directly to an action.

The goal-oriented action could be linked with what is often in the literature described as intention, and intentional action. The formation of the intention, especially the occurent one as described by Mele, could be seen as forming a goal, especially since it is characterized by its executive character. The goals can change if the expected utility changes, or another action, with a higher utility is presented, and moreover this scheme can account for action control, where a discrepancy between the stimulus and initial goal, in which case the goal is adapted or replaced by a second one. Moors describes a scenario where our goal of being respected at work is discrepant with a college making fun of us in the professional setting. The two elements, in opposition to one another, trigger a second goal, which is supposed to reduce said discrepancy, and can trigger a number of actions, from joking and walking away, to shouting, and giving up the initial goal to be respected, depending on the utility value of each of those actions. The model is further linked with Davidson's approach where:

The first goal in this cycle provides the value of one action outcome and corresponds to what philosophers call a desire. The expectancy corresponds to what philosophers call a belief. The action tendency corresponds to what philosophers call an intention or goal to act. (Moors, 2017, p.68)

The quote above is creates a clear link between the standard model, and the discussions in psychology, where the (free) action is caused by a belief/desire pair, with the addition of the context of the action, as it provides us with different possible actions.

The classic approach, where stimulus-driven and goal-directed processes are competing with one another, and the former is taken to usually win, was challenged by Moors and colleagues (2017, forthcoming). The alternative approach, following the standard model, also divides the action domain into two processes: goal-directed and stimulus-driven based on the content of representation. However, the argue that the goal-directed processes can be also automatic, range from simple to complex, and in fact often win with stimulus-driven processes (depending on the value of the goal).

Conclusions.

Where are we now: what causes the action?

In this chapter I have analyzed the various causes for intentional action. As we have seen previously:

(a) Free and intentional action overlap, and possibly all free action is intentional action (although the relationship might not hold in reverse, so not all intentional action is free). In the discussions conducted here I do treat them synonymously, as I am only interested in the intentional actions, which are at the same time free.
(b) In order to be considered intentional, the action has to be caused in a specific way.
(c) Most broadly, the action cannot be caused by only a physical brain process (e.g. the conclusions of Libet), but rather should be caused by a conscious decision.
(d) Unclear notions: what do we understand as a cause- the different possibilities were covered in the previous chapters, where several positions were identified, including classical humean approach (where causal relationships are backed up by laws, and what we observe are regularities, the position emphasized the importance of the perception of causal links), counterfactual positions (where causality is synonymous with counterfactual dependence), interventionism (where causality is understood as making a difference to the effect), and production accounts (where the cause produces the effect).
(e) Unclear notions: what counts as a cause in the 'right way of causing'?

The last problem was the main issue explored in this chapter.

What are the different possibilities?

We have seen in the discussion here that at least two main types of events or processes can be identified as causal for intentional, or free, action. The first two competing positions, marking the two major trends in philosophy of action, were anti-causalism of Anscombe, and causalism of Davidson. The latter argued that causal explanations of action cite actual causes, and intentional action is one caused by reason. Those reasons are in turn created by a combination of a desire and belief, e.g. my desire to drink something and my belief they is potable water in the tap, together create a reason for me to go to the kitchen, pour myself a glass and drink it. Anscombe on the other hand argued that reasons are not in fact causal, and our actions are caused by our intentions, a possibility initially rejected by Davidson. However in his later work, he recognized that intention cannot be reduced to a specific pair of desire and belief, and therefore forms a separate category. Intentions then became a way for us to organize our various desires, and stabilize our beliefs.

In the current discussions on free will, the majority of writers in philosophy argues for causality of intentions, dividing them into two, e.g. Mele with his proximal and distal intentions, or occurent and standing intentions, or three, e.g. Pacherie with D-intentions, P-intentions and M-intentions, categories, and they are usually (except for the occurent and standing) based on their proximity of the intention to the action. The intentions that are further from the action seem to be more abstract, connected to more general planning, and not dependent on the context of our choices e.g. forming an intention to go on holidays next year, only to some degree takes our current situation into account. On the other hand the intentions that are closer seem to be more specific, linked to a specific plan of execution of the action that takes into account the context of the action. Those are also the types of intentions that are in the scope of interest of researchers in neuroscience, who usually focus on the intentions that immediately result in action. This approach was criticized by Gallagher(2006), who argues for including in what we call intentional or free action generation not only immediate causes, but also ones that are further away, like distal intentions, or causes of the motor intentions.

Another problem that was covered at length in this chapter is the identification of an action itself. For neuroscientists it seems obvious that a simple overt movement is an example of an action, while philosophical discussions are divided, for some an action is an overt movement, while for others the movement is a consequence of an action (e.g. Alvarez), or a part of an action (e.g. Pacheire).

The proper definition of what an action is, can be essential as, when applied to the experimental domain, the various understanding of the term can potentially lead to a variety of interpretation of the experimental data. Neuroscientists currently are claiming that an action is an overt movement, and therefore intentional and free action will also be an overt movement. Said movement has a number of possible causes, and only if they are of a specific type, the movement can be free. Yet if we understand action as Alvarez does, a movement can be a result of an action and not an action, therefore the experiments provide us with no new information about an action, only about a movement. On the other hand on Pacherie's account, where movement is part of the action, the experimental results can only be understood as presenting a part of a story.

One thread that binds it all.

The one element, often left undiscussed, or taken for granted, that seems to be shared by all authors working on the causes of our (free) intentional action, is their mental nature. In most of the definitions discussed throughout this chapter, intentions and reasons were usually presented as mental, somehow connected to the physical realm. The nature of that connection is often unclear, at least in the realm of philosophy of action, however when we think of the connections between said arguments and the discussion on free will, it can become central. The problem of mental causation is probably as old as the discussion on free will and other classical problems in philosophy, and currently it has no no-fail remedy: it seems almost impossible to provide a solid account on how mental could be causal (as mental) in the physical realm. Mele in various points of his work (e.g. 2009) argues that it has to be either the mental intention or its physical realizer that causes the action for it to be intentional, and as I argue free. This approach seems to assume however, that at least sometimes it will be the mental and not the physical realizer, which will be the cause of an action. The underlying assumption here seems to be twofold: (1) there is a connection between the mental and the physical, therefore we are not dealing with a dualistic scenario where two substances are competing to cause an action, and (2) sometimes mental causes can be causal through the virtue of being mental, therefore that it causes the action as mental and not only via its physical realizer. For the latter the problem is not the existence of mental per se, but rather the biggest threat is the mental being epiphenomenal, having all its causal powers because of its physical realizer.

If we then assume that for an action to be free, it has to be caused in a specific way, which means being caused by something mental (either reasons or intentions), causally efficacious mental becomes central to free will. In the next chapter the problem of mental causation will be discussed with more detail, and it connection to the problem of free will, will be further explored.

Chapter 7. Mental causation.

The previous chapter's main aim was to investigate the types of causes that are taken to be at the source of voluntary, intentional and free action. Setting aside the implicit, and the explicit assumptions of the Libety-style experiments, I have analyzed the discussions in philosophy of action and psychology. The possibilities included:

(1) Davidson's event-causal approach, where the voluntary action is caused by a belief/desire pair, that constitutes a reason.
(2) Anscombe's intentions.
(3) Mele's proximal and distal intentions. Mele's position can be seen as an extension of the later Davidson's postion, since after initially rejecting intentions as a separate category, later Davidson changed his mind, and decided to treat them as a separate category, not reducible to reasons.
(4) Pacherie's model, where, based on the time lapse between the intention and action, but also depending on the content of the intention, classifies them onto three, rather than just two categories, depending on the relation an intention has to an action, the D-intentions, the P-intentions and the M-intentions. The reasons for this new division are threefold: (1) it seems that the intentions come after each other in a sequence, where the moment the second one begins, the first one ends, and (2) the control over the action is given primarily, and almost exclusively to the second intention, one closer to the action itself (on Mele's model, this would be the proximal intention), and (3) on the previous models the activations specific to the motor system are not included, or at least not to the extent needed. The intentions are then divided onto three categories: (1) D-intentions, or distal intentions, which seem connected to the deliberation and planning; (2) P-intentions, or proximal intentions, which are linked to the intention to act now, up to the completion of the action, and (3) M-intentions, or motor intentions, which neuroscientists call motor representations. The latter are responsible for the organization of motor sequence. The three types are expected to work in sequence (starting from the distal, through

proximal, to motor), but the transition from one to the other does not entail that one stops being active when another one appears. Therefore the forming of an M-intention does not mean that either, the D-, nor the P-intention has become non-active, assuming potential to work in parallel. Pacherie's model not only assumes that all the different intentions are at work at the same time, but also differentiates between the roles they play both in action execution, as well as in controlling of both the action, and the movement

(5) The goal-directed and the stimulus-driven actions, investigated in psychology, where the former are most likely to be the type to support free will, although not all goal-directed action will be also free.

I have hypothesized that the central assumption, both for the positions discussed in the chapter immediately preceding this one, as well as for neuroscientific approach, is the mental character of those events. Therefore the question of causality in connection to the assumption of the causal efficacy of the mental, in the generation of voluntary and free action, becomes the question about the possibility of the former as a condition of existence for the latter. In other words, if we require that the causes, that are leading to free action, are mental, we have to show that mental can be causally efficacious in causing an overt, and physical, action.

Mental causation can be seen as a sub-problem of the broader question of what is causality, but more often is discussed as a stand-alone problem, where the nature of such causation is secondary to the issue of their mutual dynamics. Or in other words, first we need to determine the relationship that holds between the mental and the physical, and then, based on the answer, evaluate if a causal link between a mental cause and a physical effect is possible at al. While most of us, if asked if our mental can influence our behavior, would without hesitation answer yes, showing how exactly that could be possible in the physical world is not straightforward. Intuitively we are convinced that our experiences, and our decision guide our daily life, from the most mundane to the biggest actions. We think that the mind not only exists, but it also does things, that have consequences in the physical world.

In this chapter I will try to cover most of the difficulties that arise when we investigate the nature of the mental, and its relationship to the physical. I will discuss the different ways to describe the relationship between them, and what the consequences of each approach are. I will then focus on the different modes in which the mental is taken to exist, which will include dualism, reduction and nonreductive physicalism. I will show why the latter is our best bet, when we want to defend free will, and be able to incorporate the results of the neuroscientific experiments. As emphasized several times already, neuroscientists frequently refer to dualism, and argue their experiments prove its falsity. While in philosophy Descartes original idea, involving a world where the mental and the physical are two independent substances and as such enter causal relationships is not the dominant one. In contemporary discussions rather different variations of nonreductive physicalism are most commonly used to describe the relationship between the mind and body.

Dualism gave way to different versions of physicalism, both reductive as well as nonreductive. For the latter the mental items, whatever they might be, are then taken to be dependent on, or as often put, realized by the physical realm. For reductive physicalists the relationship between them is even simpler: the mental is the physical and can be reduced to it, where mental properties are physical properties, or can be also eliminated (the mental is just an illusion). However, given the way experiments are constructed and conducted, complete reduction of the mental to the physical is, as I will show, unlikely. The way experiments are done implies the mental and the physical are not reducible to one another, treated specifically as competing causes of the overt action. Moreover reduction, as I will discuss with a bit more detail further, comes with its own set of problems, ones connected specifically to the mental to physical relationships, and some having to do with the laws of reduction in general.

While neuroscientists openly reject dualism, they usually do not specify which kind of dualism they have in mind, presumably assuming that dualism is synonymous with substance dualism. That is not necessarily the case, and different versions of property dualism like, specifically applied to the problem of free will agent causation (most notably by Timothy O'Connor). However, the neuroscientific assumption on the disconnect between the mind and the brain, prevalent in the experimental approach, is compatible with even the strictest substance dualism, where the mental, disconnected from the physical, influences the physical causes. Here I will argue that dualism does not present a serious threat to free will, and nonreductive positions are more likely, as they fit better into the experimental domain.

As for the structure of this chapter: I will begin by discussing two extremes: reduction and dualism. The former reduces the mental states to the physical ones, and I will argue that given multiple realization of the physical and the modern experimental approaches reduction is not likely. Moreover I will, briefly, mention other problems with reduction, such as the lack of strict laws of reduction, however this plays a less significant role in this dissertation. Dualism on the other hands assumes that the two: the mental and the physical, are two separate substances. It has been argued previously (e.g. discussed in Mele, 2014), that dualism is a threat to free will, and in this chapter I will show this is not necessarily the case. From there I will move to the position, I argue, is best suited to deal with the type of causation we have in mind in connection to mind-body interaction. I will then discuss two popular varieties of nonreductive physicalism: anomalous monism and functionalism. I will then discuss the different possible connections that link the mental and the physical, including realization, supervenience and dependence. I will also give a closer look at the notion of multiple realization.

I will then proceed to discuss what is often considered to be one of the biggest problems for mental causation, namely the Causal Exclusion argument. I will show that, based on the argument and the solutions presented to the problem of dualism, epiphenomenalism is in fact the biggest threat to the possibility of free will. I will also argue that, when applied to the neuroscientific experiments, epiphenomenalism is more problematic than reduction, as it avoids most of the problems linked with the latter, while having the same practical consequences.

The two extremes: dualism and mind-brain identity theories.

Reduction.

Dualism and mind-brain identity theories form two extremes of the discussions on the relationship between the body and the mind. The former grants the mind independent ontological existence, where it is considered distinct from the physical and as such entering causal relationships. The latter holds that the processes of mind are identical to the processes going on in the brain, can be reduced to them. This understanding of reduction does not cover, not does it attempt to, all the kinds of reductions, nor will I differentiate between reduction and reductive explanation. The literature on the topic is vast, therefore I will only limit myself to the types immediately relevant for this dissertation, namely mind-brain identity theories, token reduction and type reduction. I also discuss reduction under a specific state of the art, therefore I, at least theoretically, allow for the possibility that at some point in time, it could be possible to reduce everything to physics. Although I do find that unlikely. Therefore the type of the positions I will consider here, can probably be called physicalist rather than reductionist (since reduction is ontologically independent). The notion of reduction itself is ambiguous, but here I will use it strictly in connection to the mind-brain relationship, therefore by the concept I will mean that everything about the mental can be explained through a physical process, and that all the causal powers of the former drain to the causal powers of the latter.

Mind-brain identity theories were popularized by Feigl (1958) and Smart (1959) who, independently, proposed an argument for what is now called type physicalism. In their view the mental is nothing over and above the physical, and

[…] a vast arrangement of physical particles, but they are not, over and above this, sensations or states of consciousness. They are just behavioral facts about this vast mechanism such as that it expresses a temptation (behavior disposition) to say […] I am in pain. (Smart, 1959, p. 143)

Therefore, according to Smart, all that exists is a mechanism, a physical, ongoing process, that encompasses all sensations, beliefs and consciousness. Being in pain is just a mechanism that triggers the expression 'I am in pain'. It is unclear however, at least from this brief introduction, if the sensation of actually being in pain, and expressing I am in pain, would be an identical physical process, or rather two separate processes. It is however clear that Smart (1959, p.145) considers sensations identical to brain processes, although it does not seem to entail that the description of pain as a sensation can be just as easily translated into a description of a brain process.

These types of theories were an example of a specific trend in theories, where the higher level properties were generally seen as reduced, or reducible, to the level of basic physics. Between those mind-brain identity theory was an example of type reduction, not only the specific instances (tokens) of mental were reduced to the specific instances of the physical, but the mental as a whole was reduced to the physical as a whole. This is crucial, since, as we will see further, because of the multiply realizable nature of the mental, this type of reduction has been largely abandoned. On the other hand token identity is still used.

Another theory of reduction, often discussed in philosophy of mind, and most notably by Kim, whose work prominently focuses on the Causal Exclusion argument, as we will see further, is the one presented by Nagel (1949, 1961, 1970). He introduced so called bridge laws, which are taken to contain both the predicates of the reduced and the reducing science. Nagel defines reduction where it:

> *[…] is effected when the experimental laws of the secondary science (and it has an adequate theory, its theory as well) are shown to be the logical consequences of the theoretical assumptions (inclusive of the coordinating definitions) of the primary science.* (Nagel, 1961, p. 352).

Therefore from the two theories, or the two sciences, can be reduced to one another if the laws from the secondary science can be derived from the laws of the primary science, with the help of so-called bridge laws. However, as we can see from this brief analysis, Nagel would consider reduction as a type of explanation. The validity of the bridge laws that allow reduction has been widely disputed, with some questioning if they can be interpreted as identity statements at all (e.g. Fodor, 1974). Moreover some, e.g. Churchland (1986) has found nagelian model unacceptable and needing a complete change.

Are these models applicable to the questions at hand? Firstly, the type reduction of the sort discussed in the mind-brain identity theories has been quickly abandoned, after the development of functionalism and other nonreductive positions. Mental events are there defined through their function, and, moreover, are multiply realizable. I will investigate this notion more thoroughly further on, but, in a nutshell, it means that the same mental event, e.g. pain, can be realized by a variety of physical states, e.g. different in humans, animals, and aliens. This was the main reason for the rejection of the type level reduction of mind to brain.

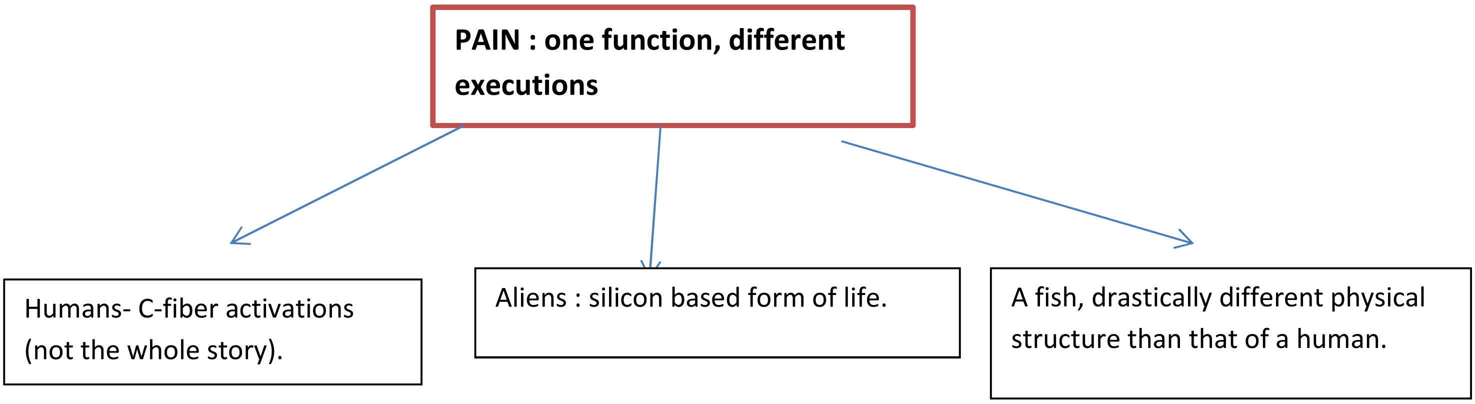


Figure 15. Multiple realization of pain.

If we look again at the neuroscientific models, such as Libet's, we will see that phenomenon such as conscious decision, or consciousness (in the work of e.g. Dehaene, 2014) are not taken to be reducible to mere physical processes. For the ontological angle everything is physical, but so is the case in nonreductive physicalism. Moreover what seems to be at the center of neuroscientific interests is not the ontological status of the mental, but rather if it can do anything, if it has an impact on the physical. Additionally many neuroscientific theories seem to focus on higher level processing abilities of the brain, and recognize multiple feedforward and feedback loops, assuming that the story here cannot be reduced to a single neuron. And if reduction is pushed to its limit, one physical process, on a very basic level, would have to be responsible for the high level processing, and what we consider free decision. Especially if the question is not about ontology, but about causality (therefore not if the mental exists but rather if it has causal powers of its own). Since I consider that unlikely, given the current state of the art knowledge, I think it is safe to say that other positions can be better suited to attempt to resolve the issues discussed in this dissertation.

Dualism.

On the other side of the spectrum of possibilities from mind-brain identity is dualism. The position was originally developed by Descartes, and treats the physical and the mental as two separate substances, which are capable of interacting causally with one another. In his Meditations Descartes describes an "I", which we would now call a mind, as completely distinct from the body (Meditation 6). Both the former and the latter are real, exist, and cannot be reduced to one another. Smart, the proponent of mind-brain identity, while opposing dualism argued that even if both, dualism and reduction, are taken to be equally possible, the principles of simplicity would push us into choosing the latter rather than the former, because it does not call for a large number of irreducible psychophysical laws, which would be *difficult to swallow*.

Dualism comes in different varieties, two most common being substance and property dualism. The first one, with its difficulty in explaining how would mental actually exist, and how it would be possible for the two substances to, causally, interact, has been largely discredited. The second variety is more plausible, and some of the positions in the free will debate could be considered examples of it. Property dualism does not claim that there are different substances, settling on arguing that there are different properties in the world. In the debate on free will the agent causation approaches are considered to be an example of dualist theories, but it is not completely clear if they would have substance dualism in mind (e.g. it is unclear if O'Connor's would not accept some form of substance dualism, with the agent being a separate substance). On those accounts, the agent is, in some sense, an independent substance and the actions of the agent are caused by the agent himself and not by the atoms that constitute such agent. This can be interpreted as substance or property dualism, where, in the latter case, the agent is token identical to the physical atoms that constitute him, but is not type identical to them.

Dualism has been implicitly suggested to be a necessary condition for free will, as some neuroscientists (e.g. Haggard, 2011; Desmurget, 2009) argue that their experiments conclusively show its falsity, and through that, the impossibility of free will. The experiments seem to assume that, in order for an action to be considered free, it needs to be caused in a specific way, by a conscious decision. The core of the argument against free will based on Libet's results comes from the temporal ordering of the two causes: the physical comes before the mental, therefore the action is not free. It is not clear if this temporal ordering can be seen as more than just that (e.g. Schlosser, 2014) but it establishes the importance of the causal source of an action as a factor in determining if it is voluntary. Moreover, since no connection between the two causes is investigated, it would suggest that they are seen as independent, suggesting a dualistic approach. Furthermore the mental is taken to arrive 'from nowhere' as not only its connection to the physical, but also its antecedents are missing, which suggests not only dualism, but an Aristotelian approach, where the conscious decision is an unmoved mover. This can be interpreted in six distinct ways: (1) the conscious decision does not have causal antecedents, which seems unlikely, (2) the decision had only mental antecedents, all of them not connected to the physical realm, (3) it has mental antecedents, all of which are connected to the physical, (4) it has both mental and physical antecedents connected to one another, (5) it has both mental and physical antecedents not connected to one another, and (6) it has only physical antecedents.

The first, second and fifth interpretation seem to suggest that the mental is a separate substance, or is in a different way independent from the physical. In the interpretations presented above I do not assume reduction of the mental to physical, for a number of reasons. Besides the classical problem, linked to multiple realizability of the mental, where the same mental event (e.g. pain) can be connected to a variety of different physical realizations (e.g. both humans and dogs can feel pain, while their physical build-up is different), it is also unlikely given the way the experiments are constructed. The assumed reduction would not only necessitate tracking of the physical basis of the conscious decision, since those two processes would be identical, it would also change the core question, from which one: the conscious decision or the physical brain process is causal, to: which one of the two causal processes is the cause of an action. The latter has much less impact on the question of free will. It also puts into question the specific distinction that is drawn between the two, and, moreover some of the scientists (e.g. Desmurget et al., 2009) argue that Libet's approach show the falsity of dualism, suggesting its existence was among the initial hypotheses, and was proven incorrect.

If we assume that, in order to have free will, our conscious decisions cannot have physical or neurological antecedents, the argument will be shaped as follows (Mele, 2012; in his reply to a problem conceptualized by e.g. Vargas):

1. *All conscious decisions have neurological antecedents.*
2. *A conscious decision is freely made (or is an exercise of free will) only if it has no neurological antecedents.*
3. *So no conscious decision is freely made (no conscious decision is an exercise of free will).* (Mele, 2012, p.426)

The argument captures two essential features at the core of neuroscientific approach: the presumed disconnect between the mental and the physical, and that the mental is also not caused by anything physical itself. Therefore a very strict version of dualism is assumed.

This assumption is however not uncontroversial. Firstly, this type of dualism is not the way most philosophers see the mind-brain relationship, which is reflected also in the discussions on the topic in philosophy of action. There almost generally accepted is their relationship. Therefore, besides the implicit assumptions of the experimental approach, nothing suggest we have to accept such a strong version of dualism, or, in fact dualism at all.

Mele (2006, 2009, 2012) argues that a simple solution can be used against such arguments, and an action can be identified as free if it is caused either by an intention or its neural realizer. Here he seems to be following the footsteps of Davidson, and argues for some form of nonreductive physicalism, with supervenient relationship between the mental and the physical. However, while such solution works against dualism, it brings another issue to the foreground, one which might not be so easily solvable.

With dualism being unlikely, and potentially incompatible with the materialistic or physicalist worldview of scientists, and reduction equally problematic, none of the two transpires to be a valid position in the debate, if we aim at a worldview in which mental cannot be completely reduced to the physical, at least in the functional sense. As I have argued previously, this type of reduction, also does not seem to be interesting for the neuroscientists.

The position that seems then most likely to account for the mental, as being, at least in the functional sense, independent from the physical, while at the same time not requiring the existence of a separate substance, is nonreductive physicalism. While the position has become standard in the debates on the connection between the mental and the physical, as we will see further is also the source of one of the major puzzles to be solved, the causal exclusion.

Nonreductive physicalism.

The quick dismissal of the mind brain identity theories was largely due to two positions, which, separately argued for the impossibility of complete reduction of the mental to the physical, although on different grounds: anomalous monism and functionalism. The former, initiated by Davidson, in line with his approach to causality in general, pointed out the lack of strict laws that would connect the mental and the physical. Putnam on the other hand pointed out the multiply realizable nature of the mental. The former was also the first one to use the concept, that now became the classic in the debate on the connection between the mental and the physical, namely supervenience, in this context.

Anomalous monism.

In the previous chapters, most notably in the one on causality, I have discussed various aspects of Davidson's approach. I have described his view on what causality is, as close to those of Hume, therefore bringing the issue of laws to the foreground. As we will see this is paramount also in his approach to mental causation. To summarize:

(1) A voluntary action for Davidson, is one caused by a belief/desire pair which together constitute a reason. Therefore in this chapter I will also focus only on his view on

propositional attitudes, such as beliefs and desires, as they are essential to the topic at hand.

(2) Causal connections between events are accessible to us through co-occurences, that are backed by laws. Unlike Hume, but closer to the approach of Anscombe, Davidson also emphasizes the importance of language, thigs are what they are because of the way they are described, which again will be central to his views on mental causation. Therefore the fuller the description of the cause, the more likely it is we will discover true causal connections.

Anomalous monism was most fully detailed in *Mental Events,* where he argued to have provided an account that supports the causality of the mental, without reducing it to the physical. At the same time he did not claim that the former was ontologically separate from the latter. Moreover Davidson argued that the mental was token identical to the physical, while at the same time not being type identical, therefore some residual reduction was assumed. This however should be seen through Davidson's broader approach to the question of the mode of existence, as it is closely connected to the language used in description of the events. Therefore, while ontologically token identical, the same did not hold on the level of description, where it mattered most. Thus descriptions of a specific cause, given in the mental terms, could not be reduced to a description of the same cause given in physical terms.

Davidson's position is of particular importance, since he introduced to the debate the notion of supervenience, which has become the standard way to think of the relationship between the brain and the mind. The article additionally triggered the debate, between Davidson and Jaegwon Kim, which led to formalization of the problem faced by all nonreductive physicalists, the Causal Exclusion Argument.

As in the case of causation in general, also for mental causation the laws that cover causal relationships are paramount. His main argument concerning the causality of mental events is based on recognizing both their causal dependence as well as their anomalousness, therefore the failure of being covered by some rigid law (here I treat this as a failure of being covered by specifically a causal law).

Anomalous monism is based on three major principles, the Principle of Causal Interaction, which:

[…] asserts that at least some mental events interact causally with physical events (Davidson, 1980, p.208).

This, very intuitive assumption, is illustrated by Davidson using an example of what causes a ship, in this case Bismarck, to sink. He argues that the person that caused the sinking (by e.g. firing the rockets) in order to do so, had to move their body in a specific manner to execute this task (e.g. press a button). This movement was, in turn, caused by a mental event or a series of mental events, like deciding, or believing that this is a good decision. The parallels with Libet-style experiments assumptions are evident here. According to Davidson two prominent examples of the mental influencing the physical can be immediately found, one is action and the other perception. Those provide best examples to illustrate the causal impact of the mental on the physical.

Davidson's second principle, Principle of the Nomological Character of Causality, states that:

[…] where there is causality, there must be a law: events related as cause and effect fall under strict deterministic law. (Davidson, 1980, p.208)

This principle is in line with the humean approach of Davidson, that was discussed a bit more in detail in the previous chapters here. Every instance that exhibits a causal connection between two events is an observation of a co-occurrence that is backed by a strict law. Importantly Davidson only seems to allow strict, and not ceteris paribus, laws.

The third principle, Anomalism of the Mental, states that:

[…] there are no strict deterministic laws on the basis of which mental events can be predicted and explained. (Davidson, 1980, p.208)

Two things can be said here. Firstly, the mental events are anomalous, therefore their impact on the physical realm is not covered by any strict laws. By those Davidson presumably means laws of physics, or something close to it. Secondly, and this bears close relation to the classic debates on free will, Davidson seems to equate here determinism, or at least deterministic laws, with predictability. Since the interactions of two events are causal, when they are covered by laws, it seems that determinism is predictability, something that e.g. compatibilists oppose.

It is hard to immediately see how the three principles would be compatible with one another, since the third one stands in opposition to the two initial ones. Davidson's view on causality is centered on the importance of laws for any account of causality, as his argument, based on Hume's regularity approach, binds causation strongly with the idea that for every causal event there is a law that covers it. However according to the third principle no such laws exist for the mental realm, and yet, based on the first principle, the mental is somehow causal. This seeming duality is closely related to what Davidson considered events, which, on his account are:

[…] taken to be unrepeatable, dated individuals such as the particular eruption of a volcano, the (first) birth or death of a person, the playing of the 1968 World Series, or the historic utterance of the word, 'You may fire when ready, Gridley'. (Davidson, 1980, p.209-210).

Therefore all the specific events that we discuss are token, where the mental is reducible to the physical, and therefore, it seems, covered by the physical laws. While on the type level, where mental is usually described, the same laws cannot, and in fact do not have to, hold, and the third principle of anomalous monism, while it entails that the mental does not fall under any strict laws, it does so as mental. This is only achievable on the type level, as on the token level the mental is identical to the physical, and any causal relationship is one between physical events.

One could ask how then is Davidson's position really any different from that of e.g. Smart. The main difference between the two is that the latter extents the commitment to the reduction up to the level of types, arguing that the mental, as type, is reducible to the physical. Davidson on the other hand is only ever committed to the token reduction. Some (e.g. Herbecke) argue that on Davidson's position

[…] that each token mental event is identical to a token physical event, where being a mental event is defined over the reference relation of a mental predicate to the event in question. (Herbecke, 2008, p.119)

The difference between the two accounts, is connected to the rejection of laws of reduction that could be used to show that mental is in fact identical to the physical. Or at least that such an identity cannot be established on the type level, while presumably possible on the token level.

The difference between the mental and the physical is mainly identifiable on the descriptive level, where physical events are described in the purely physical vocabulary, while mental events are described using mental terms. Since Davidson does not provide an extensive analysis of what would either vocabulary be, I assume he takes the folk intuitions to be sufficient here, and for the physical descriptions he would use the language of hard science.

In the case of the mental terms, Davidson is in fact only interested in propositional attitudes, and, therefore does not discuss other possible types of mental vocabulary. Therefore he refers to notions such as i.e. believing, intending, knowing etc., which are usually linked to living things rather than objects, and are accompanied by some form of 'psychological construction'. Davidson's aim does not seem to be to give an in-depth analysis of the nature of mental vocabulary, but rather to emphasize the way it can be used to explain the causality of the latter. Moreover the central role here seems to be played by the notion of intentionality (after Bretano).

The general aim of Davidson is to argue for an event monism, where on the token level the mental and the physical are identical, combined with, at the same time, property dualism, where the various properties are distinguished from one another, using their descriptions. As he himself describes it he is aiming to achieve

[…] a version of identity theory that denies that there can be strict laws connecting the mental and the physical. (Davidson, 1980, p.212)

The full spectrum of possibilities accounting for the way the mental and the physical are connected, or related to one another, can be divided into four categories: (1)

nomological monism, where the laws connecting the mental and the physical events exists, which in turn leads to their identity; (2) nomological dualism, where the laws connecting the mental and the physical exists, and the two are distinct and not reduced to one another; (3) anomalous dualism, where the laws fail and we have two substances and (4) anomalous monism. As we have seen throughout this subchapter, the latter is the position Davidson is arguing for.

Since the mental characteristic are not governed by any laws, their relationship to the physical has to be described using a different notion, and here Davidson borrowed a term used in philosophy previously but never in this context, namely supervenience. As has been mentioned throughout this dissertation, currently supervenience has become the standard for describing the non-causal dependence between the mental and the physical, and as we will see further it comes in a variety of strengths. However initially, in Davidson's theory, it was only mentioned almost in passing, where he stated that the mental characteristics are;

[…] in some sense dependent, or supervenient, on physical characteristics.(Davidson, 1980, p.214).

He then defines the relationship as one that:

[…] might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect. Dependence or supervenience of this kind does not entail reducibility through law or definition: if it did, we could reduce moral properties to descriptive, and this there is a good reason to believe cannot be done; and we might be able to reduce truth in a formal system to syntactical properties, and this we know cannot in general be done. (Davidson, 1980, p.214)

Supervenience described in this manner captures what is not considered the core of the notion: the co-changing nature of the mental to physical connection (when an event changes with respect to its mental properties it also changes with respect to its physical properties), and its co-dependence, where two events or objects identical in all physical respects cannot differ in the mental ones.

Therefore, to summarize what has been said so far: the mental is dependent on the physical, can be a cause of other physical occurrences, but such relationship is not covered by strict laws. Why are the three principles not contradictory then? Davidson argues that relationships such as causality or identity are normally relations that hold between events independently from the way they are described. Therefore the same laws should hold regardless how we describe events. However, the differentiation between the mental and the physical is done on the basis of description. Therefore, in spite of the existence of causal links between the mental and physical, since they hold between individual events, but described in mental terms, this cannot be covered by laws, since the description of those events can change. Aside from the different descriptions, ontologically all the events are physical, as it is still monism, therefore for some Davdison's arguments are mainly epistemological. However, since the language and the way things are describe is crucial, and central for Davidson, there is no major difference between ontology and epistemology here. The fact that his position is monistic one means that:

[…] not only can every mental event be uniquely singled our using only physical concepts, but since the number of events that falls under each mental predicate may, for all we know, be finite, there may well exist a physical open sentence coextensive with each mental predicate […]. Indeed, even if finitude is not assumed, there seems no compelling reason to deny that there could be coextensive predicates, one mental and one physical. (Davidson, 1980, pp.215-216)

The irreducibility is then, similarly to the way mental and physical are distinguished from one another, not ontological but rather nomological in nature. However, since the events are not type-reducible, only token-reducible, it seems that no strict laws can be found to cover general, true statements, only particular true statements. While strict laws are not possible, lawlike constructions are permissible, which have forms of laws but are not laws. These statements also support counterfactuals, which, as we will see further, is an important feature for most solutions for the Causal Exclusion argument.

Based on different generalizations, Davidson distinguishes two different types of laws: (1) homonomic, which are those that give us reasons to believe adding different conditions in the same language could become a finished law; and (2) heteronominc, which are generalizations which make us believe that there is a law at work, but this law can be stated only by switching to a different vocabulary (i.e. mental rather than physical). Strict laws, on Davidson's account, are only homonomic and not heteronomic, since the latter lack the strictness of a law. Again, the mode of description is essential here, since events described in a physical vocabulary with generalizations and laws expressed in the same vocabulary are, or have the potential to be, actual laws. When the vocabularies differ between the description of events and generalizations, the events cannot be covered by laws, or the heteronomic generalizations cannot be turned into laws. The reasons for this difference are discussed by Kim (1985), where he interprets Davidson's main argument as one where:

The mental system has a certain essential characteristic X and the physical system a certain essential characteristic Y, where X and Y are mutually incompatible. Laws linking the two systems, if they exist, would 'transmit' these characteristics from one system to the other, leading to incoherence. Therefore, there can be no laws connecting the mental and the physical so long as the two systems are to retain their distinctive identities. (Kim, 1993, p.201)

Therefore, laws between the mental and the physical would require reduction, and since such is impossible on the type level, the conditions for existence of laws cannot be meet, because the descriptions cannot be reduced.

How does this all connect to the free will problem and the causal source of action? Firstly, anomalous monism was designed to account for the mental events' causal powers, which would allow for the type of action generation required for free will. Davdson clearly argued that:

The anomalism of the mental is thus a necessary condition for viewing action as autonomous. (Davidson, 1980, p.225)

While not all autonomous action is free, and the two are not synonyms, it seems that all free action is autonomous, therefore they can be treated as such in the context of this dissertation. However, two problems emerge. Firstly, as argued by Kim (1985), by holding that the links between mental and physical do not have nomological character, Davidson could achieve the opposite and make free agency impossible. On anomalous monism sentence of the form 'All people with a mental property M also have a physical property P' would not constitute a law. On both Davidson's and Kim's accounts however there is an intimate relation between laws and causality, which can influence the truth and falsity of such statements. Therefore the lack of such relationship might lead to falsification of the causality of the mental.

To summarize Davidson's position:

(1) The mental is taken to be causally efficacious, while at the same time, being anomalous.
(2) No strict laws connect the physical and the mental domain, since those have to hold regardless of how the events are described, which is crucial in the mental and physical distinction.
(3) While strict laws are impossible, generalizations and ceteris paribus connections can hold.
(4) Mental is reducible to the physical on the token level, however not on the type level. Therefore individual events can be seen as purely physical, however type mental (as beliefs, in general) cannot. It seems also that on the token level all events are subjected to laws, which could potentially give priority to the physical over the mental.

While Davidson aims to reject reduction and provide an explanation on how the mental could be a cause of at least some physical events. However his approach might ultimately lead to mental, while not reduced to the physical, to be epiphenomenal, therefore lacking any causal powers. Since there are no laws that would connect the physical and the mental, Davidson's position seems to give priority to the physical explanation, as this one is governed by strict laws.

Throughout this chapter I will hypothesize, and explain why, I think that epiphenomenalism might be the biggest threat to free will, if our focus is on the manner in which an action is caused. Anomalous monism can be seen to support epiphenomenalism, and if so it will inevitably fail to produce a satisfactory account of free will. Davidson himself argues (e.g. in *Thinking Causes*) that most of his work, especially in philosophy of action, was aimed at providing an account where mental is the cause of our behavior. However he might be falling short from his goal, not because on his account mental could not be causal through the physical, and through the supervenience connection. Rather because he fails to show how mental could be causal qua being mental (or the infamous qua problem).

Another serious shortcoming of Davidson's theory is its lack of clarity on what exactly constitutes a strict law. It is clear that it should hold regardless of how the event is described, which excludes the mental, but it might be just as difficult to find physical laws that would hold in all possible circumstances.

Probably the most important Davidson's contribution to the mind-body debate is introducing the notion of supervenience to it. Since then the rudimental definition grew to include a number of varieties, with different, counterfactual strength.

Functionalism.

The other popular option among nonreductive one, currently almost a standard in the literature on mind-brain relationship, is functionalism, popularized by, among others, Putman (1967) and Fodor (1968, 1974).

Functionalism is potentially compatible with both dualism and reduction (although Putnam specifically argues against it), however, over the recent decades, it has become increasing more attractive for physicalists, who claim that those different states are most likely physical, and for nonreductive physicalists, who take multiple realizability to grant the possibility that the mental can be causal. The thesis of functionalism extends beyond the mind-body problem, and it is used by some to that reduction of any theories, also special science ones, to the physical level, might be impossible (e.g. Fodor, 1974). For physicalism, functionalism unlike the mind-brain identity thesis, allows for creatures other than humans to experience the same sensations, a serious shortcoming for the predeceasing theory, while at the same time keeping its core assumption: that causality happens on the level of the physical.

Putnam, in *The nature of mental states* starts from the analysis of the well-known example used by the brain-mind identity theorists, where pain is identified as a brain state. He argues that, for this to be true, the two have to be in some sense synonymous. He wants to argue that the two do not fulfill this condition, and pain

*[…] is not a brain state, in the sense of a physical-chemical state of the brain (or even of the whole nervous system), but another kind of a state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism. (*Putnam, 1975, p.433)

The state of being in pain, or any other mental property, is a functional property, where its internal construction is not essential, instead the way it functions, and the role it plays in the complete system is emphasized. For instance, being in pain is not the matter having a certain internal state, or at least not only, but being in the state of believing that there is something wrong with our body, making certain sounds etc., and as such can be shared not only by humans but by all animals, that are capable of those states. This last point is essential here, as was already illustrated earlier in this chapter, also on the example of pain, since functionalism allows that pain as such can be experienced by a creature that does not possess C-fibers, and has a drastically different construction, not comparable to that of a human, e.g. alien life made of silicone. Following this logic, in principle all mental states can be experienced by different life forms, and are supported by a variety of physical realizations.

The basic argument for functionalism, using the example of pain was summarized by Shoemaker (2013), who introduces the problem by asking what would in fact be the best evidence that a mental state is identical with a particular physical state, like, in the case of pain: the C-fibers firing. This, in turn, would lead to the potential discovery of a universal correlation between the two states, the mental and the physical, which would lead to the conclusion that pain is present if and only if C-fibers are present and firing. The approach is compatible also with a dualist worldview, in which case the mental, while being a distinct substance, is always correlated with the physical. In the example of pain, being in this state would require C-fibers firing, which would then be correlated with a specific mental state of being in pain, which would in turn be distinct, and independent from the physical activations (of the fibers).

As we have seen in the previous chapters, there seems to be ample neuroscientific evidence that the mental, as functionally different from the physical at least exists, and cannot be reduced to a very basic physical level (e.g. one specific neuron). Dualism, is unlikely, and moreover assuming it could lead to falsification of free action. Rather than equating mental with the physical, or assuming independent substances at play, another possibility is to consider that both, the physical and the mental, have the same cause, e.g. in the case of C-fibers firing and being in pain, it could be a cut. In this case, when for both mental and physical event, the causes and the effects are identical, their causal impact is identical. Yet, since the same mental event can be identified across species, and the c-fibers firing can be found only in humans, it could commit us to stating that an animal, when cut, does not feel pain. And that seems wrong. The argument can be extended, and we could argue that pain could be realized by different physical realizations inside the same species.

The identity thesis, where the mental is reduced to something physical fails, but the problem of the manner in which we can define functional states remains. Here two ways can be identified (Herbecke, 2008), where (1) the functional states are either defined over the causal relations or (2) through the relationship they have with other states.

The first approach allows us to define functional states, roughly, as mediators between inputs and outputs of the system. In the case of pain, we could see a cut on an arm as an input and crying as an output. The feeling of pain would then be identified through the presence of those two states. In the case of relations with other, functional, states are taken not to be independent, and therefore are seen as realized by some other states (the notion of realization will be defined with more detail further). Functional states defined in this manner seem ubiquitous in the world, and were a new way of approaching the relationship between mind and body.

Human mind is the seat of a variety of cognitive capacities, including, among other, perception, language and attention, all relevant to how we usually think of the decision making processes, and through that of free will. Besides being non causal, on the mind-brain identity theory, they were also identical to some physical brain states, and unique to humans. Therefore we could be speaking of human perception, human language, human attention, and as in the example discussed previously, human pain. Perception or attention, as realized in other species would have to be redefined, as other processes, realized by a different physical, or physio-chemical, properties. Functionalism, through multiple realizability, shows that the same mental phenomenon, when defined through its function in the organism, can support the same state though a variety of physical set-ups. Therefore different mental state can be realized by different physical states. Applying any of the nonreductive positions to the free will problem, requires however further explanation of their core notions, which will be provided below.

Multiple realizability thesis.

In the literature we can find a number of variations of the multiple realizability definition, with different levels of strength. Those different approaches aim at accounting for (1) what does it mean that the physical states that realize the mental states differ, or in other words, how different can they be, (2) does the relationship holds only between the types, or also between tokens.

For the first problem, broadly defined, the concept allows one functional state to have a different realized in every situation it can be found it, which supports the idea that every state can be realized by a multitude of different realizer states, as is in case of pain across different species. More specific definition expressing different strength can be found, for instance, in Polger (2002) and Sahpiro (2004), who distinguish four varieties of multiple realizability thesis, abbreviated as MRT:

(1) *Weak MRT: At least some creatures that are not exactly like us in their physical composition can have minds like ours*
(2) *SETI MRT: Some creatures that are significantly different from us in their physical composition can have minds like ours.*
(3) *Standard MRT: Systems of indefinitely (perhaps infinitely) many physical compositions can have minds like ours.*

(4) *Radical MRT: Any (every) suitably organized system, regardless of its physical composition, can have minds like ours.* (Shapiro, 2004, p.7)

While all of the approaches above could be applicable to the mind-body problem, and in fact were designed to do so, for the scope of this dissertation even the weakest form of the definition would be sufficient. In case of free will, what we need to argue against reduction of the mental to the physical, as this is the core task of the multiple realization thesis here, it is sufficient to state that all creatures that are not exactly like us have a mind. In fact, even a weaker version would be sufficient: where creatures exactly like us, meaning other humans, can have the mental event be realized by a variety of different states. As we will see further on, for the solutions to the Causal Exclusion argument, the multiple realization is employed in connection to counterfactual analysis, where the same mental state, in the same individual, is taken to be realized by a variety of different physical states. Therefore not only C-fibers in general, but, e.g. a specific C-fibre, different each time, would realize the property of being in pain. Moreover, in connection to the debate central here, it is essential to properly define what we want 'mind' to be. One of the possible angles of the free will discussion is the question if free will is unique to humans, or if we can assume that also animals have free will. It seems that, if we think that mental causation is essential to the possibility of free will, applying other possibilities proposed by Shapiro, would lead to leaving the door open for the free animals. Of course other components could make it impossible, however, if the question would be taken purely from the perspective of the causes of our actions, and the causal efficacy of the mental, assuming that radically different systems can have a mind of their own, leaves this possibility open. While I remain agnostic on the topic, my intuitions guide me to believe that free will, as defined here, is uniquely human, and while it might be wrong, I will limit myself to discussing human agents. Therefore in the scope of this dissertation, when I mention multiple realization, I mean it in the weakest possible sense.

For the second problem, if it holds between types only, or includes also tokens, the examples discussed by the founders of functionalism, among others Putnam and Fodor, suggest that their aim is to describe a relationship that holds between different types (such as pain, in general). On the concrete, token, level multiple realizability seems to be absent, and a specific mental event is then realized by a specific physical state.

Some, e.g. Yablo (1992) argue that multiple realization holds also in case of tokens, and across possible worlds. The specific pain is then taken to be multiply realizable by different physical states, in humans, in different possible worlds, by using e.g. slightly different neuronal connections. The distinction is relevant for the mental causation debate for several reasons, as, when understood in terms of types it only blocks the reduction of mental to physical on the conceptual level, while when understood in terms of tokens, it has the potential to also block the ontological reduction in specific examples (e.g. Herbecke, 2008).

Multiple realization often can be seen as the core argument for nonreductive physicalism, almost its fundament. There are however those (e.g. Shapiro) who argue that it should not be taken to play such a substantial role. For Shapiro it seems that, if the two do not differ significantly in causal roles they play, and if we agree that for a neuron, the only work contributed to a specific mental state is the propagation of the chemical or electrical signal, it does not matter if that role is played by said neuron or a different structure. The contribution they both bring to the realization of the mental state is the same, therefore the mental state is not multiply realized. This would only be the case if a different realization property would bring different causally relevant properties, than for example neuron brings in our decision, to refer to the free will case. I think this specific approach however is closely connected to interventionist reply to the causal exclusion argument, that will be discussed further, where the causal contribution brought by each of the components of the causal chain is essential. Others (e.g Shea, 2013) argue that, given the neural evidence, multiple realization is not all that powerful, since the studies show:

[…] the same content on different occasions has a relatively consistent effect on the local distribution of oxygaenated hemoglobin, which is in turn coupled to local differences in firing rates (Mukamel et al. 2005). (Shea, 2013, p.107)

However, for multiple realization to be effective, it does not seem that the different realizations of the same mental property need to be radically different. The identity thesis does not specify how deep the reduction would go, would we accept patterns of neuronal firings, or do we aim at the level of particular neurons. However, it seems to me that the main threat is having the mental be reduced to a specific neuron, not that the patterns of firing are similar. Therefore I will use the hypothesis of mental being multiply realizable as still standing.

Supervenience.

Another term commonly used, usually in connection to the multiple realization, is supervenience, which is used to explain the **how** of the mental to physical connection. The application of the term is not limited to the mind-brain debate, but describes various types of dependencies. As I have discussed above, supervenience was introduced to the debate by Davidson, with a very broad definition. Since then several varieties of supervenience have been identified.

In its broadest sense supervenience commits us to minimal physicalism, the minimal commitment we have to make in order to remain in the realm of physicalist positions. In those cases, supervenience understood modally, can support reduction, as mental is taken to be wholly dependent and determined by the physical. On the other hand, supervenience can be interpreted to support multiple realization, where the two systems identical with respect to all their physical properties, are also identical with respect to their mental properties. Because supervenience is non-symmetric (neither symmetric or asymmetric) two systems identical with respect to their mental properties, can differ with respect to their physical properties. Therefore, the mental can be, in two identical systems, realized by different physical properties. Also here the multiple realization can be understood as a very weak thesis, as it describes systems that are identical, or very similar.

Supervenience entails then the co-changeability of two variables, and as Lewis argues:

[…] we have supervenience when [and only when] there could be no difference of one sort without differences of another sort. (Lewis, 1986, p.14)

More specific, yet equally vague definition can be found in McLaughlin (1995), who describes it as:

[…] the idea of dependent-variation where the dependency is of a purely modal sort. Variation in A-respects depends on variation in B-respects in that A-respects cannot vary without variation in B-respects. Variation in the supervenient A-respects requires variation in the subvenient B-respect. A-respects fail to supervene on B-respects when and only when A-respect can very independently of variations in B-respects.[…]Supervenience, so understood, is a purely modal dependent-variation relation that is reflexive, transitive, and neither symmetrical nor asymmetrical. (McLaughlin, 1995,p.18)

While describing it as modal relationship, McLaughlin does not however describe the modal force with which the relationship holds, and which possible worlds will be considered relevant for the analysis. In the discussions that can be found in the literature, the most relevant here will be those immediately connected to the problem of causal efficacy of the mental.

The discussion on supervenience, the force with which it holds, and its consequences to the mind's causal powers were probably most extensively covered by Jaegwon Kim. He has also provided an in-depth analysis of the premises of nonreductive physicalism, and showed that, taken together, they are incoherent.

In his early work Kim (1984, 1987) divided the varieties of supervenience onto strong, weak and global, and in (1988) added multiple domains supervenience, and this differentiation has remained valid, although the precise definitions has been modified. In this chapter I will focus on weak and strong supervenience, as those are specifically applied to the mind-brain relationship, since those claims are specifically about individuals, while global supervenience is used in contexts of sets or whole possible worlds.

Similarly the multiple domains supervenience does not apply in the context of this work, as the properties, or events described here are assumed to be possessed by one individual. Although global supervenience allows for including the external conditions of the individual, it is too weak to yield really interesting claims (for more discussion cf. Herbecke, 2008).

Weak and strong supervenience.

Weak supervenience is a relationship that holds when:

A weakly supervenes on B if an only if necessarily for any x and y if x and y share all properties in B then x and y share all properties in A- that is, indiscernibility with respect to B entails indiscernibility with respect to A. (Kim, 1984, p.158)

As Kim (1987) argues, this type of supervenience is too weak to properly capture the relationship that holds between different properties, including determination and dependency, often used to describe the mind-brain relationship, parallel to supervenience itself.

Conversely, the notion of strong supervenience should be better equipped for such tasks. Strong supervenience is defined as follows:

Necessarily, for any object x and any property F in A, if x has F, then there exists a property G in B that x has G, and necessarily if any y has G it has F. (Kim, 1987, p.316)

In order to distinguish between the weak and strong types of supervenience Kim often uses quantification over possible worlds, and both types hold not only within one world but between worlds as well. The possible worlds, in Kim's understanding, are all worlds possible logically, causally or nomologically, and therefore supervenience can be characterizes as modal dependent variation between two variables. In the quote above the quantification over possible worlds was replaced by the notion of necessity, but the definitions are equivalent, as for Kim necessity can replace, or be replaced by universal quantification over possible worlds (McLaughlin, 1995).

For the purpose of this dissertation, the most relevant, and formulate later, is the definition of strong supervenience, applied specifically to the relationship between mental and physical properties. It reads as follows:

Mental properties supervene on physical properties, in that necessarily, for any mental property M, if anything has M at the time t, there exists a physical base (or subvenient) *property P such that it has P at t, and necessarily anything that has P at a time has M at that time.* (Kim, 1998, p.9)

Which for Kim is equivalent to:

Mental properties supervene on physical properties, in that necessarily any two things (in the same or different possible world) indiscernible in all physical properties are indiscernible in mental respects. (Kim, 1998, p.10)

These two definitions, although taken by Kim to be equivalent, seem to emphasize different aspects of the relationship. In the first one the focus seems to be that all mental aspects have a physical supervenient basis, eliminating any "free", unbound mental events. As a consequence, this definition is not compatible with dualism, since any mental events that are not connected to physical events, possibly independent from the physical realm are eliminated. The second definition, on the other hand, shows that all things identical to one another in physical aspects, will be identical in mental aspects, which suggests the asymmetrical relationship between the two: the mental supervenes on the physical but the physical does not supervene on the mental. This opens the way for supervenience relationship to support the multiple realizability of the mental, where the latter can supervene on different physical events.

One of the problems I have discussed throughout this dissertation so far, has its source in the experimental omission of the connection between the mental and the physical. The neuronal activation that could be seen as the supervenient basis, e.g. a conscious decision would appear simultaneously with the decision, is not investigated in the approach such as Libet's, as those operate, as I argued previously, under the implicit assumption of (property) dualism. As I have argued here, a better way to explain causal efficacy of the mental, while at the same time incorporating the results of the experimental approach, without them excluding free will a priori, is given by nonreductive physicalism. The former would reject them a priori since, if we assume that the mental does not have physical antecedents, just by conducting an experiment our assumption seems to be that our hypothesis is false. Additionally the way the experiments are conducted does not in fact investigate the antecedents of the conscious decision, only the antecedents of the movement. Nonreductive physicalism, on the other hand, when applied to the free will problem, requires that we investigate both the mental conscious decision, as well as its physical supervenient basis. Readiness potential cannot be it, as the two are not synchronous, as two variables connected via supervenience usually are. Therefore, Schlosser (2014) might be right that all that Libet-style experiments are establishing, is time sequence of occurrences, or readiness potential can be in fact identified as a random brain noise, not necessarily related to the action-generating processes (e.g. Schurger et al., 2012).

Realization vs. supervenience.

Both supervenience and realizability of the mental are often discussed in a similar context, very closely to one another, which yields the question: what are the differences between the two, and are we not talking about the same thing. The problem becomes more pressing when we consider that, although in the discussions on the relation between mind and body, the notion of multiple realizability has become part of the canon, the sole concept of what it is for the mental to be realized by the physical is very rarely explained (problem is recognized by e.g. Polger, 2004; Heil, 2011; Baysan, 2015).

It is clear that the term ‘realization’ is used to describe some form of the dependence relationship, and it seems that it is mostly used as means to explain the intra-level relationships (where often the mental events are seen as higher-level properties than the physical ones). But, the same can be said about supervenience as context in which we are using the term includes the same events, or properties (mental and physical) in a specific type of relationship. Moreover, both supervenience and realization are often characterized as being a dependence relationship. Therefore, when both are defined as such, it might be hard to see clear differences between realization and supervenience.

The one major characteristic of supervenience that is missing in the notion realization is its modality. Through this inherent modality, supervenience can be interpreted, and in this dissertation will be, as referring to, or a property of, an existing relationship between the, in this case, mental and the physical. In other words, when we classify a relationship between the two as supervenience, what we are actually saying is that there is a link between them, such that its consequence is the co-changing of the two. In the case of the mental and the physical, the former is realized by the latter, and the property of that realization relationship is that the mental supervenes on the latter. Therefore, what seems to be the intuition here is that the mental and the physical are somehow connected, and a property or a feature of this connection is supervenience.

Said existing connection can be identified in a number of ways including (Heli, 2011): (1) A and B having a common cause, (2) A being B, and (3) B being a determinate of the determinable A. Therefore another connection, other than just supervenience is assumed, in order for the latter to hold.

For some (e.g. Shoemaker, 2007, 2013) realization is a better fit, a more appropriate way to describe the relationship that holds between the mental and the physical than supervenience. Better because for the latter to support physicalism, we need to assume that the relationship between the supervenience base and the supervenient property is constitutive and not causal, as has been widely done (this detail, as we will see forward, is central to interventionist approach to the topic, and their solution to the Causal Exclusion problem). Accepting this turns the supervenient base into a realizer of the property that supervenes on them (the author does not clarify this further). Shoemaker not only offers one of the few exhaustive accounts of what can be understood as realization relation, but also distinguishes two types of it: (1) property realization, and (2) microrealization.

Property realization is, according to Shoemaker, what philosophers have in mind when they use the term in the context of mind-body problem and physical realization of the mental properties where:

[…]property realization, is what we have already met with-it is what we have in the claim that a property like having C-fiber firing going on in one can realize the property of being in pain but cannot be identical with it. We speak of this as a relation between properties, but at bottom it is a relation between instantiations of properties, or what I will sometimes refer to as property instances. (Shoemaker, 2013, p.39)

In this case one property is realized by another property, and in the case of mental to physical, the mental is realized by the physical. He distinguishes it from what he calls the microrealization which is defined as:

A given type of microphysical states of affairs will be such that members of that type are apt to cause, or contribute to causing, microphysical states of affairs or certain other types, and will be caused by microphysical states of affairs of certain types. My account of microphysical realization says that a microphysical state of affairs realizes an instance of a particular property just in case the microphysical state of affairs belongs to a type whose causal profile matches in a certain way either the causal profile of the property whose instance it realizes or the causal profile of a property realizer of that property, and the microphysical state of affairs is embedded in the career of the subject of the property instance and is simultaneous with it. (Shoemaker, 2013, p.43)

In this dissertation when referring to realization I will have in mind the first understanding rather than the second, as I think it fits better the context of the specific debate this dissertation is placed in.

Another account of realization applicable to the debate in the context of the free will problem is given by Pereboom (2002), who thinks of realization as a relationship that holds between property instances. Pereboom argues for a version of nonreductive physicalism, where functionalism is rejected, and the mental states can be realized, or potentially reduced to

[…] a structure more abstract than any specific neural structure, and one that can potentially be realized by a silicon-based system.(Pereboom, 2002, p.23)

Pereboom rejects functionalism, and argues that mental states, in order to be causal, have to have structural intrinsic properties, that are also different from any other intrinsic properties that are not mental. However, since Pereboom specifically thinks that free will is not possible and the concept is incoherent (as I discussed in the initial chapters here), I am not convinced that his approach would serve well here.

Conclusions.

In this chapter I have analyzed different approaches to the question of the causal efficacy of the mental, with the problem of free will in mind.

I have argued that dualism and reduction are, for different reasons, not applicable, if we want to defend a notion of free will that is compatible with the neuroscientific findings.

(1) Dualism, especially substance dualism, is not compatible with the findings, and difficult to account for also from the philosophical perspective. Moreover most philosophers do not find the existence of neural activations that are related to the action generation, and mental events, threatening to the existence of free will.
(2) Reduction seems incompatible with the way experiments are conducted. There are no good ways to reduce the mental to the physical, although this could depend on how strict we define the notion.
(3) Nonreductive physicalism seems best fitted to explain the relationship between the mental and the physical, and moreover has been generally accepted for a while now. It is based on two notions: supervenience and multiple realization, however, as we will see further, when applied to the free will problem can have the same practical results as reduction.

I have also analyzed and defined the two core notions for the nonreductive physicalist position: supervenience and multiple realization. I have opted for a very weak version of multiple realization, to limit the scope of my discussion to the mind of other humans. I have then described supervenience as a property of a relation that holds between the mental and the physical, with a modal force.

In the following chapter I will describe the biggest threat to mental causation in nonreductive physicalism, the Causal Exclusion argument, I its consequences to the free will debate.

Chapter Eight. Causal Exclusion, and the threat of epiphenomenalism.

Where are we now?

The previous chapters have established that, in most cases, both in the neuroscientific account of causation, as well as discussions in philosophy of action, the potentially free, intentional or voluntary acts, are taken to be caused by either one of a number of mental component. In the case of the experimental approach it was the conscious decision, in philosophy of action, the possibilities range from reasons, through intentions and mental imagery.

Such assumptions led me to believe that causally efficacious mental is one of the necessary conditions for the possibility of free will, however it is not its sufficient condition, if we want to frame the problem in terms of necessary and sufficient conditions. I have previously discussed various possible ways to think about causality, and, in some cases (e.g. Mackie's INUS conditions) this is not necessary. As we have seen above, and will again see in this chapter, the way we think about causation has consequences to the possibility of the mental being causally efficacious.

I have also discussed the various way we can approach the problem of mental causation, and the status of the mental as connected (or not) to the physical. I have proposed three approaches: (1) reduction, (2) dualism, and (3) nonreductive physicalism, and I have argued that the first two are, given both the various problems each of them is facing, as well as the experimental approach to the issue, unlikely. Following this, I have then focused on the third approach, nonreductive physicalism, and I have discussed the two most prominent sub-approaches: (1) anomalous monism, and (2) functionalism. Since it is usually concluded that the former does not show how mental could be caused *qua* mental, which, in the case of the free will problem is necessary (and I will talk more about this in the coming chapters), functionalism would be better suited to account for the type of mental causation that seems desired in the free will case. In the current chapter I will focus on the biggest threat to the causally efficacious mental, namely the exclusion, most prominently covered by Jaegwon Kim (e.g. 1998, 2005, 2009).

Causal Exclusion- the argument.

Causal exclusion argument was originally aimed against the different versions of nonreductive physicalism, and specifically against anomalous monism. In his discussions with Davidson, Kim (e.g. 1985, 1998, 2005) showed that the four assumptions, shared by all nonreductive physicalists, are incoherent, and lead to, at best, the negation of the causal impact of the mental. In the worst case, for the nonreductive physicalist, they could lead to a complete reduction of the mental to the physical (although here the usual problems of reduction would still apply). Since then it has been shown (e.g. Block, 2003) that the argument generalizes to all relationships between entities connected through non-causal dependencies, like e.g. chemistry and physics. Therefore finding a solution to the Causal Exclusion seems to be increasingly more pressing, and, as I will see further, far from being achieved.

The four premises of the causal exclusion argument, implicitly shared by all nonreductive physicalists, are (1) principle of mental causation, (2) principle of irreducibility, (3) principle of causal closure, and (4) no-overdetermination principle. They can be described as:

(1) Some physical effects have mental causes (the principle of mental causation)

(2) Mental causes are distinct from physical causes (the principle of irreducibility)

(3) If a physical event has a cause at t, it has a sufficient physical cause at t (the principle of causal completeness of the physical or the causal closure of the physical domain, Kim, 2005)

(4) Events cannot have more than one sufficient cause occurring at a given time (the exclusion principle)

The first principle, or premise, reiterates the assumption, central to the nonreductive physicalists, that the mental can have causal impact on the physical realm. It states nothing about the exact manner in which this happens, and is therefore compatible also with dualism.

The second premise is a rejection of reduction of the mental to the physical realm. Also here, since the principles of such, and the connection between the two is not further explored, the premise is compatible with both nonreductive physicalism, as well as with dualism.

The third premise, described by Kim in *Physicalism or something near enough,* states that:

If a physical event has a cause at t, then it has a physical cause at t (Kim, 2005, p.15, p.43)

Therefore, the principle, while making it necessary for an event to, if it has a cause at all, have a physical cause, but states nothing about how causality is understood here, nor about the physical being the only cause. The premise does however limit the variety of causes, since it rejects the idea that only mental, without the physical, could be a cause. This however is not a very strict requirement, as it remains compatible with dualism: it could be the case that an event has two, independent causes, the physical and the mental.

The three premises so far are neither incoherent, nor do they exclude the mental from having causal impact on the physical, and doing so as mental. But, the addition of the fourth principle, of causal exclusion, changes things considerably. The premise reads:

If an event e has a sufficient cause c at t, no event at t distinct from c can be a cause of e (unless this is a genuine case of causal overdetermination). (Kim, 2005, p.17),

and as such, connected with the previous premises, while not rejecting the existence of the mental nor reducing it to the physical, shows that it simply cannot be shown to be causally efficacious.

This principle does not even refer specifically to the realm of the mental, but rather states that each event has only one sufficient cause, while at the same time not claiming that such event has to be either mental or physical. However, when combined with the premise of the causal closure it entails that there can be only one, physical cause. Kim offers a second, more detailed version of this principle, the principle of determinative/generative exclusion, which reads:

If the occurrence of an event e, or an instantiation of a property P, is determined/generated by an event c- causally or otherwise- then e's occurrence is not determined/generated by any event wholly distinct from or independent of c- unless this is a genuine case of overdetermination. (Kim, 2005, p.17).

In this version Kim, aside from limiting causes to just one, also offers an explanation of what he means by causality, namely determination or generation of one event by another. Kim claims that his approach to causation is close to the one argued for by Elisabeth Anscombe, described in the chapter on causation of this dissertation, but it has been long argued that Kim's idea of causality is that of production (without necessarily committing to the Salmon and Dowe's approach).

The premises can be then, together show one of two things: (1) the inner incoherence of all nonreductive approaches based on them, or (2) the causal inertness of the mental. Therefore, in order to show how mental could potentially be cause, most likely approach is to give up one of the premises, and keep the other three. The question which one however, has no easy answer.

What does the argument mean in practice? It is best illustrated on the well- known diagram, where the M's are the mental and the P's the physical events. The single, green arrows are used to mark causation, while the double, purple arrows mark supervenience.

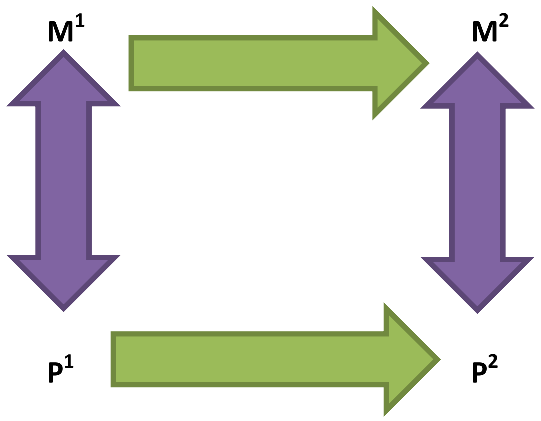


Figure 16. Kim's diagram.

The problem presented in the diagram can be interpreted two-fold, namely is it possible that the mental event M^1 is the cause of the physical event P^2, but also is it possible that the mental event M^1 is the cause of another mental event M^2. Those two questions can be illustrated as:

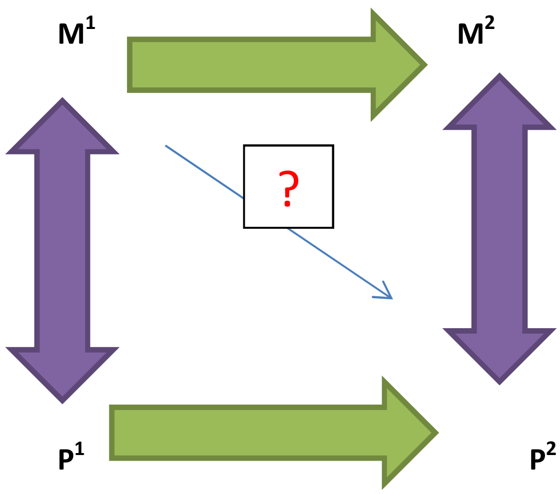


Figure 17. Kim's diagram. Mental to Physical causation.

for the first question, if there is a causal relationship between the mental and the physical, and as:

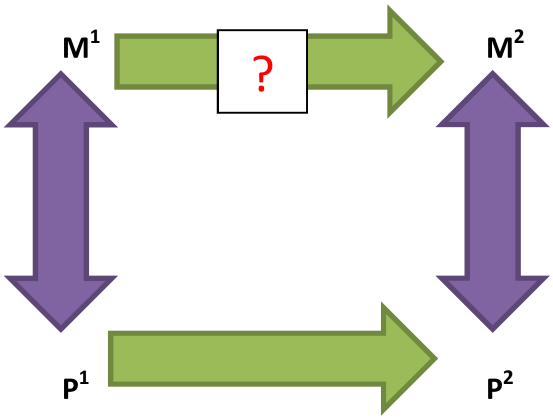


Figure 18. Kim's diagram. Mental to Mental causation.

for the question of mental to mental causality.

In the example illustrated above M^1 has a supervenient basis P^1 to which it is not reducible (based on the premise of irreducibility of the mental), while the mental event M^2 supervenes on the physical event P^2. The relationship of the latter is of course the same as the one connecting M^1 and P^1. The physical event P^2 is assumed to have a cause, and, if it has a cause, has a sufficient physical cause (based on the principle of the completeness of the physical), which in this case is P^1. Since P^2 does not have more than one cause (based on the principle of causal exclusion), all the required causal work for P^2 appears to be done by P^1. Therefore accepting the four premises of the Causal Exclusion argument, forces us to also accept that P^2 is exclusively caused by P^1 and the mental plays no role in its generation or production.

Since P^2 and M^2 are connected through supervenience, the appearance of P^2 is a guarantee for the appearance of M^2, since two individuals identical in all physical aspects have to be identical also in the mental aspects. Because, as we established before P^1 is the de facto cause of P^2, therefore P^1 is also sufficient for the occurrence of M^2, since they necessarily co-occur. This leads us to the conclusion that M^1 has no causal influence on either P^2 or M^2, and is in fact causally superfluous. The mental still exists, and is not reducible to the physical, but is causally irrelevant or epiphenomenal.

We can apply this to the example of Libet's experiment (under the assumption that readiness potential can be used as the supervenient basis of the conscious decision, although, as I have shown here previously, this is not the case), although here only the mental to physical causal impact will be questioned:

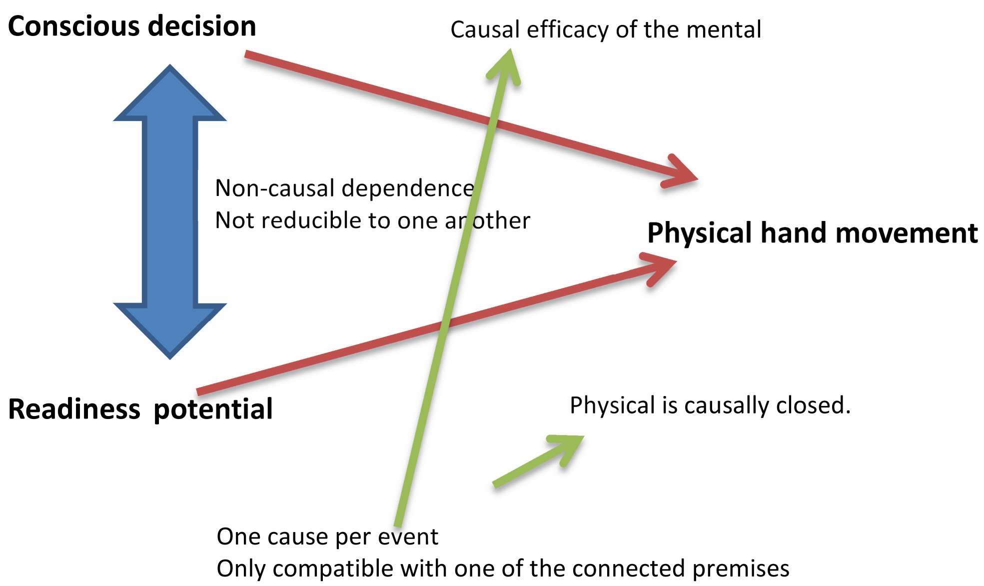


Figure 19. Causal Exclusion and Libet.

On this example we can clearly see the problem with nonreductive physicalism, as it seems to, in the light of Causal Exclusion, again support the most radical conclusions of the experiments: there is no free will. Even if we assume that the mental is neither a separate substance, as neuroscientists seem to claim, nor, reduced to the physical, and devise an account that allows for the accommodation of the experimental results, the conclusion might yet again be, that the physical movement investigated, can only be caused by the physical brain activation.

More formal representation of the threat can be found in List and Menzies (2017), who provide a revised version of the Argument specifically applied to the free will problem, where:

(1) An agent's action is free only if it is caused by the agent's mental states.

(2) Any effect that has a cause has a sufficient physical cause occurring at the same time (physicalism).

(2*) Causation implies causal sufficiency

(3) The agent's mental states are not identical to any physical states, but supervene on underlying physical states.

(4) If an effect has a sufficient physical cause C, it does not have any cause C* distinct from C occurring at the same time (except overdetermination).

The first premise links the causal efficacy of the mental, with the free action. In stating that the action is only free when it is caused by the mental, the authors express the intuitions present, as I have argued, throughout the literature in both neuroscience, as well as philosophy of action. They do not however try to put mental causation as a sufficient condition of the existence of free will, but, as I am doing here, only argue for its necessity.

Since the experimenters seem to claim that, because it is the unconscious brain activation and not the conscious mental decision that lies at the core of the action, we are not free, it is reasonable to claim that they would accept this assumption. The conclusion of Libet himself, that the conscious decision would have the veto power over the action initiated unconsciously and already under way, also relies on the possibility that the, presumably, mental conscious decision has some causal efficacy.

The second premise, referred to as causal closure, or the premise of physicalism, and states that if an event has a cause at all, it has a sufficient physical cause. The premise by itself does not exclude the existence of other potential causes, making it, according to the authors, a very weak version of causal closure, as it leaves the possibility of other sufficient causes open. List and Menzies introduce a second version of the premise, where physical closure is not necessary, but causation is constrained by the notion of causal sufficiency.

The third premise is a rejection of mental to physical reduction, and it establishes what type of relationship holds between the two.

The last premise rejects the possibility of multiple causes of one event. Accepting all four implies that mental events cannot be causes of our actions and therefore free will does not exist. Similarly to the original formulation of the Argument, it does not imply that mental events do not exist.

The core of List and Menzies approach is to show that: (1) the Causal Exclusion can be avoided on the difference making approach of causality, and (2) that the scientific experiments cannot be said to conclusively say anything about free will. The second aim depends heavily on the first one, and I will spend more time explaining it in the following chapter. However, to present the initial assumptions, as it is important here, List and Menzies argue that the difference making approach is the reason for which the last premise of the argument should be rejected, therefore the no-overdetermination argument. This is also heavily influenced by their woodwardian approach to what can be defined as causality, and, moreover if causation implies sufficiency. The latter seems to be the assumption of Kim, and therefore also List and Menzies try to apply it, in their consideration if the second premise of the article can be rejected.

This approach is not completely new, as different attempts to avoid the Causal Exclusion, thought rejection of one of the principles have been taken in the past. Those lead to: (1) reduction, if we argue that mental is not irreducible to the physical, (2) dualism, if we reject supervenience, or any other non-causal dependencies that might connect them, (3) epiphenomenalism, which keeps the mental as independent from the physical, but denies it causal efficacy, and (4) we argue that the mental, in some sense, causally overdetermines the physical. Some of them, similarly to Mackie, argue that assuming just one cause for any event, cannot be substantiated, as most of the examples we can think of, require a set of conditions to be fulfilled. The approaches involving the rejection of the principle of no-overdetermination, has been currently gaining popularity.

What is the main threat to free will?

Most of the discussions, as was repeated here several times already, focus on the threat the deterministic nature of the universe presents to the possibility of free will. The possible threat of mental causation is often omitted. This is, in my opinion caused, by focusing on the threat of dualism, and the existence of the physical antecedents of the overt movement as ta threat to free will. An approach, linked directly to the implicit assumptions the experiments that, in turn, influence the way the experimental results are interpreted. The temporal order, where the physical readiness potential appears before the mental conscious decision, is then taken to mean that the former rather than the latter is the cause. Moreover the two are seen as largely independent, therefore suggesting dualism. This, in turn, would mean that not only because the physical is first, but also because if can be identified to have a connection with the overt action, means that the mental cannot be causal here. Or, in other words, if the mental were causal, the connection between the physical and the movement would not be present. These assumptions are represented on the graph below:

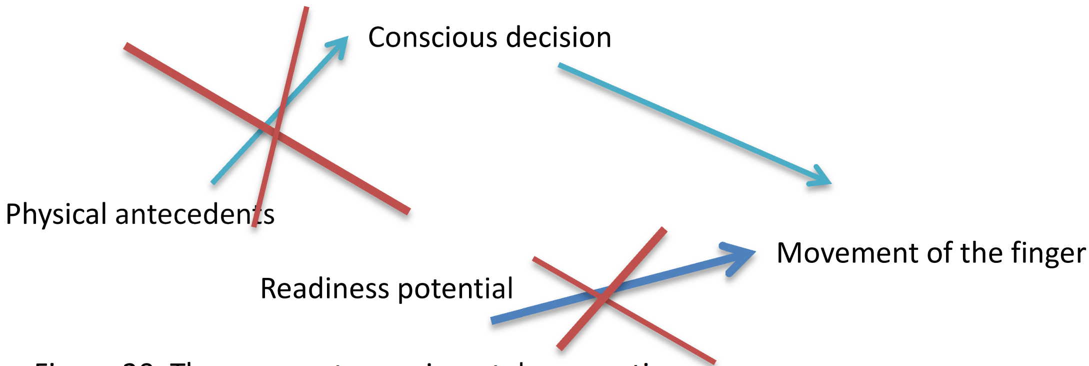


Figure 20. The apparent experimental assumptions.

As we can see above, and was discussed in the previous chapters here, the core assumptions regarding free action, that seem to be influencing both the construction of the experiments, as well as their interpretation are that: (1) for a movement to be free it has to be caused by, only, the conscious decision, (2) the decision has no relation to the physical realm, (3) there should be no connection between the physical and the overt movement, and (4) the mental should have no physical antecedents. These four assumptions are not generally applicable, to all experimental approaches to free will, and vary between positions, but the above described situation could be taken to be the worst-case scenario of assumptions. Furthermore the mental is taken, in this example, to arrive 'from nowhere' as not only its connection to the physical, but also its antecedents are missing, which suggests not only dualism, but an Aristotelian approach, where the conscious decision is an unmoved mover. This can be interpreted in six distinct ways: (1) the conscious decision does not have causal antecedents, which seems unlikely, (2) the decision had only mental antecedents, all of them not connected to the physical realm, (3) it has mental antecedents, all of which are connected to the physical, (4) it has both mental and physical antecedents connected to one another, (5) it has both mental and physical antecedents not connected to one another, and (6) it has only physical antecedents.

The first, second and fifth interpretation seem to suggest that the mental is a separate substance, or is in a different way independent from the physical. In the interpretations presented above I do not assume reduction of the mental to physical, for a number of reasons. Besides the classical problem, linked to multiple realizability of the mental, where the same mental event (e.g. pain) can be connected to a variety of different physical realizations (e.g. both humans and dogs can feel pain, while their physical build-up is different), it is also unlikely given the way the experiments are constructed. The assumed reduction would not only necessitate tracking of the physical basis of the conscious decision, since those two processes would be identical, it would also change the core question, from which one: the conscious decision or the physical brain process is causal, to: which one of the two causal processes is the cause of an action. The latter has much less impact on the question of free will. It also puts into question the specific distinction that is drawn between the two, and, moreover some of the scientists (e.g. Desmurget et al., 2009) argue that Libet's approach show the falsity of dualism, suggesting its existence was among the initial hypotheses, and was proven incorrect.

If we take mental causation to be necessary for free will, and moreover this is possible only in a dualistic world, the argument against it could look like this:

4. *All conscious decisions have neurological antecedents.*
5. *A conscious decision is freely made (or is an exercise of free will) only if it has no neurological antecedents.*
6. *So no conscious decision is freely made (no conscious decision is an exercise of free will).* (Mele, 2012, p.426)[18]

The argument captures two essential features at the core of neuroscientific approach: the presumed disconnect between the mental and the physical, and that the mental is also not caused by anything physical itself. Therefore a very strict version of dualism is assumed.

This assumption is however not uncontroversial. Firstly, this type of dualism is not the way most philosophers see the mind-brain relationship, which is reflected also in the discussions on the topic in philosophy of action. There almost generally accepted is their relationship. Therefore, besides the implicit assumptions of the experimental approach, nothing suggest we have to accept such a strong version of dualism, or, in fact dualism at all.

Mele (2006, 2009, 2012) argues that a simple solution can be used against such arguments, and an action can be identified as free if it is caused either by an intention or its neural realizer. Here he seems to be following the footsteps of Davidson, and argues for

[18] This specific formulation comes from Mele's article, however it is important to note it is based on an argument put forward by Vargas and Fisher.

some form of nonreductive physicalism, with supervenient relationship between the mental and the physical. However, while such solution works against dualism, it brings another issue to the foreground, one which might not be so easily solvable.

Mele's approach works well to show that the underlying assumption of dualism found in experiments on voluntary action, is not a major threat to the possibility of free will. He shows that the connection between the mental and the physical, without reducing the former to the latter, has long been assumed in philosophy, and therefore there are reasons to be suspicious when holding that dualism is necessary here. However, if we aim to show that the mental can be causally efficacious qua mental, Mele's argument fails to do that. And this is in fact the main issue stemming from the causal exclusion argument.

The center of Mele's account is the assumption that, in causal generation of voluntary action, it is **either** the physical/neural realizer **or** the mental intention that is the cause of it. This only holds if we assume that, at least in some cases, it is the mental and **not** its physical/neural realizer that can be identified as a genuine cause. If that is not the case, the position is no longer, practically, disjunctive, as it would mean that it is always the physical realizer that does the entire causal job. And therefore, if we think, and as I argued here this is a very widespread view, that for an action to be voluntary/ intentional/ free it has to be caused 'in the right way', which involves the mental intention to be the cause, Mele's position fails to show this, which can be illustrated as proposed below:

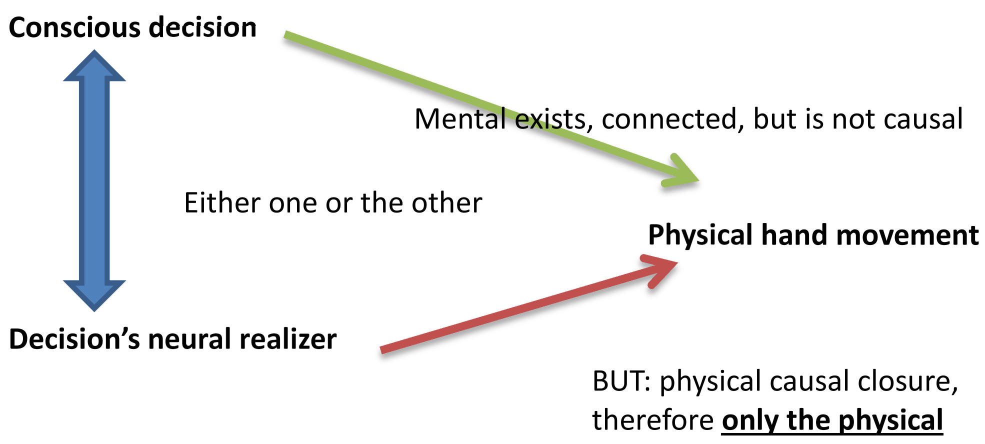


Figure 21. Mele's solution and Causal Exclusion.

As we can see in the graph above, since Mele, allows that it is, on his account sometimes, the physical that is causal, given causal closure, he unintentional falls into epiphenomenalism, where the mental remains distinct from the physical, but has no causal impact on the action generation. Therefore the link marked by the green line, is not causal. The way Mele aims to save mental action initiation, mainly from the threat of dualism, can be seen as closely related to, or even an example of, nonreductive physicalism, where the mental is dependent, or connected, to the physical but neither reduced to it, nor a different substance. This exact type of positions was challenged by causal exclusion, and therefore is sensitive to the conclusions of List and Menzies. While Mele's position is not a full account of mental causation, the phrasing of the solution suggests that those are the underlying intuitions. If it is the case the causal exclusion argument applies to Mele's position as for any other nonreductive one, except from dualism, which, as was shown here, is not a major threat to free will. Therefore, while trying to show the mental can be causal, and linked to the physical, in causal processes resulting in an action, Mele's position can be seen only the connection between the two, however fails to account how the former could be causal, given the exclusion problem. Therefore, while assuring that reduction and dualism are not big issues, Mele's argument not only does not work against the threat of epiphenomenalism, but entails the latter.

In its practical application epiphenomenalism is, for the causal relationship between the mental and the physical realm, worse than reduction. On the reductive approach we can still hold that the mental was causal, as what is threatened by the reductionists is not the causal efficacy of the mental, but its irreducibility to the physical. Therefore the mental would be, at the same time, causal and identical with physical. Of course this does not allow for the mental being causal qua mental, yet it is causal nonetheless. For reduction we can illustrate the situation as follows:

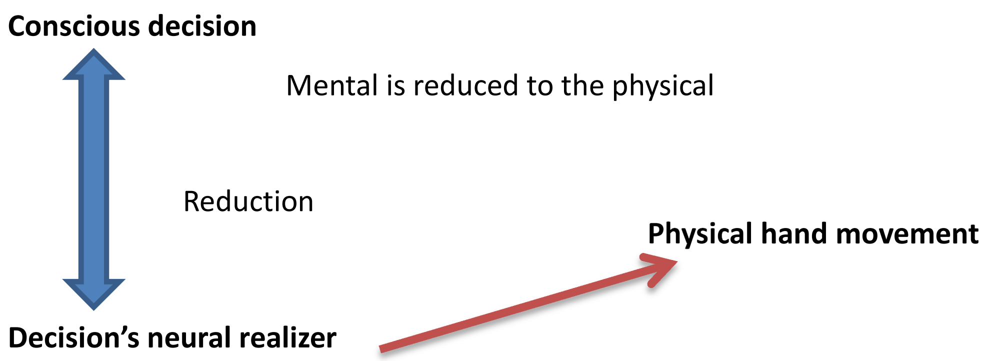


Figure 22. Mental reduced to the physical.

Here, in comparison to the picture where the mental is not reduced, the green line which signified the connection mental could have with the physical, is missing. However the red line, marking causality, is the same in both pictures. Therefore, while there are significant differences between the two pictures when we consider the irreducibility of the mental, those differences seem to disappear when we consider causality independently from the irreducibility. While on reductive approaches, the mental is causal (as physical), on epiphenomenal approaches the mental is not causal. On both accounts the mental is causal only though its physical realizer, which is insufficient to argue for the possibility of free will, which in turn, requires that the mental be causal qua mental. This is due the two core assumptions I have discussed here at length: that for an action to be free it has to be caused in the ‘right way’, and for this to happen, the cause has to be mental. When we think about the practical application of these two assumptions epiphenomenalism is no different than reduction as both of them lead to the mental not being causal, and the physical doing all the causal work. The two positions differ in the reason why the mental cannot be the cause, for reduction it is because the latter can be fully reduced to the physical. For epiphenomenalism it is because it has not causal powers other than those of its physical realizer, therefore, only the physical realizer can be considered a cause, in the sense necessary for free will. Yet on both of them mental is not causal, and therefore free will is not possible.

This failure of Mele’s position often is not fully recognized. Even if mental causation is mentioned, often the focus is again on deterministic mental causation, going back to the problem of determinism, and deterministic agents (e.g. Deery and Nahmias, forthcoming). On the other hand the dangers linked to the possibility of mental causation as such, are sometimes completely dismissed. However, as was shown above, Mele's position, because of the causal exclusion challenge, cannot be taken to dismiss the issue.

Between reduction, dualism and epiphenomenalism it is the latter that presents the biggest problem then. The first two positions can be shown to have serious problems, which epiphenomenalism avoids and moreover the latter is compatible with most of the assumptions about the nature of the mental, and allows us to retain the agent's conscious experience, the basis of how we think of ourselves as agents. And yet, it yields the same results for the causal efficacy of the mental as reductive approaches.

Does neuroscience establish that mental is epiphenomenal?

Do Libet's experiments show that mental is epiphenomenal? There has been a lot of discussion on the issue, also in the span of this dissertation, although the arguments usually revolve around causality of events, rather than possibility of mental causation. Some (e.g. Schlosser, 2014) argue that none of the experiments conducted using Libet's paradigm actually establishes causation, but rather either temporal order (i.e. Libet, 1982, 1085; Soon et al., 2008), or temporal co-variation (i.e. Haggards and Eimer, 1999). The fact that the chain of events follows from an unconscious physical event, through the conscious decision, up to the action, does not entail that it is the brain activation rather than the decision was the causal origin of the overt movement. Schlosser argues that it is the lack of controlled interventions that yields this result, however, this assumption relies on a very specific understanding of causality, i.e. intervention and not production.

Others argue that the experiment does not cover the whole process that ultimately led to the action, as the intention was formed when the subjects received an instruction from the experimenter. The decision that took place during the experiment itself was induced by those instructions. There is a number of other authors (e.g. Sinnott-Armstrong and Nadel, 2011) that expressed similar conclusions, that Libet's experiments do not in fact show that the decision is not the cause of an action, because they do not exclusively show that the cause of the movement is in fact the physical rather than the mental event. Those arguments, while questioning what is causal, implicitly assume that both potential causes, can be causes, or, that the mental can be causal as mental, something that the Causal Exclusion argument explicitly refers to.

The Causal Exclusion could be considered a non-threat to this specific neuroscientific approach, as in case of Libet-style experiments, they do not account for, nor aim to explain how the mental could be related to the physical. In the experiment the supervenient basis of the conscious decision is not investigated, and it is also possible that another mental event supervenes on the readiness potential, not being reported by the individual. Therefore, while questioning the causality of the mental, the experiments can be seen as a contribution to the mental causation debate only if we accept the possibility of dualism. Only then the mental can be investigated without simultaneously keeping track of the physical, since, if we accept nonreductive physicalism, the two are connected. However the lack of connection brings about the question if the conclusions of the experiment can really support a conclusion that it is the physical and not the mental that's the actual cause.

Where does this leave us?

In the past few chapters I have argued that:

(1) Free will problem is usually discussed in connection to determinism, either causal determinism of the universe, or the deterministic agent.

(2) The neuroscientists however, in their experimental approach, assume that, for an action to be free, it has to be caused in a specific way, namely through the conscious decision causing it.

(3) I have then identified that similar intuitions are also present in the philosophy of action, where a voluntary action, for which free will is a subset of, and in the scope of this dissertation treated as synonymous, is seen to be caused by either: an intention, a reason, or mental imagery.

(4) I have then argued that all those possibilities share one, crucial characteristics, they are all taken to be mental events.

(5) I have then analyzed what does it mean to cause something, and presented a number of options as to how a connection between the two events can be seen as causal.
(6) Then I have proceeded to the specific type of causality, namely mental causation.

In this chapter I have identified the biggest problems for free will, when we take mental causation to be central, focusing mainly on the causal exclusion argument.

Things that have been argued in this chapter:

(1) The way we approach causality can have consequences to the types of answers we obtain to the exclusion problem: the latter is usually formed with the production account in mind, while using an interventionist account is largely argued to support mental causation against the exclusion argument.
(2) There are three major possibilities to think about the relationship between the mental and the physical: (a) dualism, (b) reduction, (c) nonreductive physicalism. I have shown that the last one leaves the most room to accommodate the experimental results, while at the same time ascribing the mental some causal role.
(3) But, nonreductive physicalism, both anomalous monism and functionalism, are threatened by the causal exclusion account, where their premises are inconsistent and the positions are not able to show how the mental could be causal qua mental.
(4) Failure to do so leads to epiphenomenalism, which, as I have shown, in practice has the same consequences as reduction.

Chapter Nine. The already explored ways to move forward.

This dissertation, up to this point, had one major goal: to establish that mental causation should be seen as one of the necessary conditions for the existence of free will, and explaining how free will could be possible requires an account of mental efficacy. Therefore the assumption here is that the mental has to be causally efficacious for free will to be possible, however it does not guarantee its existence, therefore is only a necessary and not a sufficient condition. Other reasons why free will might still not be possible, even if we show that mental can in fact be causally efficacious, include the usually deterministic nature of the universe, or the existence of alternative possibilities for us to choose from, all of which were covered in the first chapters here. Deterministic mental causation might still leave no room for what we usually consider free will to be, deterministic universe where possibilities do not exist might be just as problematic. Moreover the problems such as the issue of luck, and the problem of disappearing agent, both discussed in the previous chapters of this dissertation, remain (with the last one possibly becoming even more pressing, given the solution I want to sketch here. Therefore establishing the causal efficacy of the mental is just a first, small step, in the free will endeavor, but, as I argue here, a necessary, while often ignored, step.

The following chapter will explore the possibilities to move forward, when we have established that a working account of causal efficacy of the mental, it might be necessary, for establishing how free will could be possible, to show the way out from the claws of the Causal Exclusion. Since this chapter also serves as the first from the conclusive ones, I find it necessary to, again, revisit the main arguments from each of the chapters.

Short summary of the previous chapters.

The philosophy and the neuroscience of free will.

This dissertation follows in fact two, major, problems in philosophy, namely the problem of free will and the questions grouped under an umbrella of causality, which include such as what constitutes a cause, what is causality and if causal explanations cite actual causes of actions.

In the first chapter I have analyzed the classical discussions in the free will debate, focusing on the compatibility of free will with determinism, as well as the problem of existence of alternative possibilities the agent can choose from. For the former, I have identified the various possible approaches (compatibilism, libertarianism and hard determinism), while for the latter I have presented the, now classical, argument against the issue presented by Frankfurt. I have also discussed some of the results of experiments, conducted by philosophers, exploring the folk intuitions on free will and its compatibility with the state of the universe. The major positions in the discussion are (1) libertarianism, where free will is incompatible with the deterministic universe, (2) compatibilism, where free will is compatible with determinism, and (3) hard determinism: free will is impossible (either because the universe is deterministic, or because the concept is in itself inconsistent).

It should be noted that incompatibilists usually do not make a definite statement on the actual nature of the universe, settling on pointing out the incompatibility of determinism and free will. The same can be said about compatibilists, who argue that free will is possible even if the universe is deterministic. Libertarians on the other hand tend to argue that universe is in fact indeterministic, and therefore the possibility of free will is preserved. Since I argue that the problem of mental causation holds regardless of the nature of the universe, I have chosen to remain agnostic on the topic.

I have additionally covered another classic problem in the free will debates, namely the so-called principle of alternate possibilities, which, as a condition of possibility for free will, treats the existence of different potential courses of action, for the agent to choose from. In the classical rebuttal of the problem Frankfurt has claimed that the possibilities do not have to exist actually, what matters are the reasons for which the action was performed. The non-existence of actual choice is a factor only when the individual performed an action **only** because he/she could not have done otherwise. As long as an agent recognizes his own reasons for action, he can still be held responsible for them.

Besides these two major lines of discussion, I have also analyzed the experimental data pointing to the importance of belief in free will, and the impact it can have on the human behavior. I have also decided to treat free will as synonymous with (1) moral responsibility, and, as was argued in later chapters (2) voluntary action, as free will seems to be a subset of such, therefore each free action is presumably voluntary, although the reverse does not hold. I have recognized that the two might not in fact be synonymous, but rather intertwined, however, the problem of mental causation, if in fact relevant, touches all of them equally.

In the second chapter I have turned to the empirical data, used in the discussions on free will to prove, or disprove, its existence. I have focused mainly on the Libet-style experiments, as they seem to have become a standard in the neuroscientific approach to the problem, with their variations being still executed. Besides analyzing the conclusions reached by the various investigators from the data gathered, I have also pointed out the various reasons why those interpretations rely on some initial, often implicit assumptions. Those, in turn can in fact heavily influence both the construction, as well as the interpretation of the data, and ultimately, the conclusions that can be drawn from them.

The most significant of those is the assumption of an existing disconnect between the physical and the mental domain. In the experiment the conscious decision is taken to arrive seemingly from nowhere, not connected to any tracked, ongoing or preceding brain activations, and its neural correlated are not investigated, which provides a strong basis to suspect a dualistic interpretation at play. In fact many neuroscientists openly argue that their position shows the falsity of dualism, without acknowledging other possible options, including reduction and nonreductive physicalism. I have argued throughout that the last one is particularly important, as it allows for incorporation of the experiments into the philosophical discussions without ad hoc rejecting the mental as the possible cause.

What do we talk about, when we discuss causality in connection to the problem of free will?

Because experiments focus on the causality of actions, and their interpretation of the results is linked to the notions such as cause, and causality, I have focused my next chapter on what we understand, or can understand, but those notions.

The positions I have covered defined causality as: observable co-occurrences backed by laws, counterfactual dependence, production, and manipulation. In the first two cases, the general connective feature is the reliance of some form of law, most likely inaccessible to us, to justify and solidify the connection between the two events. In the case of two last approaches, they represent a turn from the law-oriented approach, where causality is backed by some form of connection, either through transfer of something, a mechanism, or, in the last case, making a difference. Each of the interpretations of what a causal connection between two events can consist of, has consequences to the mental causation debate, and in turn to the free will problem.

To further substantiate my argument, that the causal efficacy of the mental is an essential part of the free will problem, I have then analyzed what kind of causes are taken to appropriately cause an action that can be defined as free. Such actions are usually seen as caused by: (1) belief/desire pair that forms a reason for acting (Davidson), (2) intention (Mele and Pacherie), or (3) conscious decision (Libet). I have argued that one characteristic they all share, is their mental nature. Moreover, in most cases, the mental and the physical realm are seen as connected, and the only exceptions seem to be the arguments by neuroscientists (at least in the discussions on how the voluntary action is caused). Therefore, in order for the 'right way' of causing to be possible, causal efficacy of the mental seems to be required.

Which problems appear when mental causation is seen as necessary for the existence of free will?

If we accept that mental causation is central for the existence of the possibility of free will, we need to firstly, decide what kind of connection holds between the mental and the physical and what the features of this connection are. Secondly, we need to face the dreaded Causal Exclusion. The latter shows that the four principles of nonreductive physicalism, taken together, are inconsistent. When applied to the free will problem however, shows that without mental causation, free will is impossible.

The classical formulation of the argument, most famously discussed by Kim (e.g. 1998, 2005), was adapted to the free will discussion by List and Menzies (forthcoming), where the premises read as follows:

(5) An agent's action is free only if it is caused by the agent's mental states.

(6) Any effect that has a cause has a sufficient physical cause occurring at the same time (physicalism).

(2*) Causation implies causal sufficiency

(7) The agent's mental states are not identical to any physical states, but supervene on underlying physical states.

(8) If an effect has a sufficient physical cause C, it does not have any cause C* distinct from C occurring at the same time (except overdetermination).

The argument can be then illustrated through the interaction of various possible causes and effects, where the green arrows are used for hypothesized causal links, purple arrows for supervenience, or other non-causal dependencies, and the blue arrow for the mental to physical causation. The question marks show which causal links are problematic:

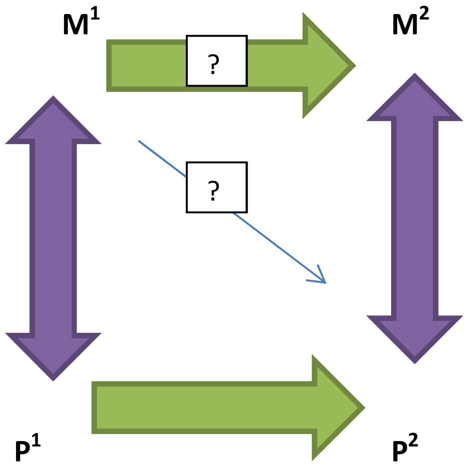


Figure 23. The problem with mental causation.

To briefly summarize the problem again, as was done in the previous chapter, M^1, which signifies a mental event, has a supervenient basis P^1, which signifies a physical event. The former does not reduce to the latter (based on the premise of irreducibility of the mental). The mental event M^2 supervenes on the physical event P^2. Here I am using supervenience as a relationship that connects them since this is the very broadly accepted hypothesis. As we have seen in the previous chapter, this can be also a relation of realization or another form of dependence. The relationship connecting both pairs in the example is identical. The physical event P^2 is assumed to have a cause, and, if it has a cause, has a sufficient physical cause (based on the principle of the completeness of the physical), which in this case is P^1. Since P^2 does not have more than one cause (based on the principle of causal exclusion), all the required causal work for P^2 appears to be done by P^1. Therefore accepting the four premises of the Causal Exclusion argument, forces us to also accept that P^2 is exclusively caused by P^1 and the mental plays no role in its generation or production.

Since P^2 and M^2 are connected through supervenience, the appearance of P^2 is a guarantee for the appearance of M^2, since two individuals identical in all physical aspects have to be identical also in the mental aspects. Because, as we established before P^1 is the de facto cause of P^2, therefore P^1 is also sufficient for the occurrence of M^2, since they necessarily co-occur. This leads us to the conclusion that M^1 has no causal influence on either P^2 or M^2, and is in fact causally superfluous. The mental still exists, and is not reducible to the physical, but is causally irrelevant or epiphenomenal. As I have mentioned before, this is not a major issue for some, however, when applied to the question of the causes of our actions, and especially to the free will problem, epiphenomenalism seems to differ little from reduction.

The conclusive chapters of this dissertation will firstly, here, cover the commonly discussed solution to the causal exclusion problem, and point out their shortcomings. The latter will be discussed with the specificities of the free will problem, and with the experimental approach in mind. Therefore I will show that what we should want is an account that allows for the causal efficacy of the mental as a necessary, but not a sufficient, condition for free will. And one that can incorporate the findings of neuroscientific experiments without those immediately requiring either unpopular ontology, or leading to an immediate falsification of free will, based on mostly implicitly and not necessarily correct assumptions. The latter will be the job of the conclusive chapter of this dissertation, which will show my own solution to this problem, and discuss the potential applications, and the ways it can incorporate the experimental results into the philosophical discussions on free will.

Is Causal Exclusion the end of the free will dream?

Over the years there has been a number of attempts to either solve, or avoid the Causal Exclusion problem altogether. Most commonly the solutions focus on rejecting one of the premises, and they can be divided into four categories where: (1) mental causes are reduced to the physical ones, which requires an account of laws governing such reduction, (2) mental is independent from the physical, which means accepting dualism, entailing in turn the rejection of supervenience relationship and the connection between the mental and the physical; (3) mental causes are epiphenomenal, and the causal influence they might have comes from their neuronal realizers (although they are still recognized as distinct from them, and the two are linked through the usual supervenience relationship), and (4) mental causes overdetermine the physical causes, therefore rejecting the no-overdetermination principle (while the mental still remains realized, or supervenient on, the physical). Throughout this dissertation I have shown that, when applied to the free will problem, some of the solutions are not available. Dualism not only is controversial, which by itself would not be a major issue, but also leads to the potential impossibility of free will, and does not allow for a smooth incorporation of the scientific evidence. Reduction is unlikely both because of the way experiments are conducted, as they suggest that the mental is, somehow, different from the physical. Moreover multiple realization remains an issue for all reductive accounts, even if we only decide that it has a limited scope (e.g. Shea, 2013). Epiphenomenalism on the other hand, as I have shown is the biggest threat for a working account of free will, because, while it has the same practical consequences of reduction, it also accommodates the scientific evidence. Therefore, as I have argued previously, epiphenomenalism can be seen as the biggest threat to free will. The rejection of the last, no-overdetermination, principle can be a stand-alone solution, where the specifics of what causation means are not necessarily discovered, and combined with, or justified by, introducing another concept of causality, another way to avoid the exclusion problem.

Another way to avoid the Causal Exclusion problem is to change the way we define causation, since the argument was created with a very specific approach in mind, namely a production account. This path was chosen by the proponents of Woodward (initially Shapiro and Sober, Raataikainen, and only later Woodward himself), who claim that applying interventionist account shows that the mental, and not the physical, is in fact the cause of an action. Those solutions usually, also, require that one of the premises is falsified, but it is because of the changed notion of causality, not because we want to solve the Casual Exclusion per se.

Different approach to causality: Interventionism.

The problem of Causal Exclusion itself was constructed with a specific understanding of causality in mind. While Kim sometimes argues that his approach to the topic is Anscombian, this is unlikely, as the descriptions seem to play a limited role for Kim, while being central for Anscombe. Instead, one of the production approaches seems to fit much better to Kim's expectation of what constitutes a causal relationship. Therefore, we could argue that the argument was constructed with a specific understanding of causality in mind, and changing how we understand the notion, will also have consequences to the argument. This was the intuition that guided interventionists to apply their interpretation of what constitutes a causal link, to the mind-brain relationship.

With the successful applications of interventionism in different domains, some authors (e.g. Shapiro and Sober 2007; Raatikainen 2010; Woodward 2015; List and Menzies 2017) became convinced that interventionism can also solve the notorious Causal Exclusion Argument and account for causally efficacious mental processes. Shapiro and Sober (2007) specifically argue that Woodward's ideas provide the means through which the micro-effects of macro-causes can be identified experimentally, leaving room for an account of mental causation. The experimental component is of particular interest here, as this would make interventionism attractive also for the neuroscientists working on the question (and as we have seen before on the rare occasions when the notion of causality is discussed in neuroscience, they lean towards the interventionist account).

According to Woodward (2015), the approach to causality he is advocating (most exhaustively in Woodward, 2003) allows for a much more natural way to capture the relationship between the mental and the physical, seen as a supervenience relationship where macroscopic properties are realized by the microscopic ones. Again, it is important to note here that causality, for interventionists, is nothing over and above making a difference. This stands in opposition to the classical approach, where some form of physical connection, or exchange of something between the two variables, are seen as necessary.

The interventionist solution involves the core idea of manipulating the mental and the physical, and through that establishing what the actual cause is. In the first glimpse the proposed explanation is appealingly easy. All that is needed for M_1 to be a cause of another event, like M_2, or a behavior, is the existence of a possible intervention that can be performed on M_1, such that, while changing its value, also the value, or the probability distribution of M_2, will change as a consequence.

To recall the main definitions, the direct cause is defined as (Woodward 2003, p.59):

A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V.

Contributing cause on the other hand is defined as:

A necessary and sufficient condition for X to be a (type-level) contributing cause of Y with respect to variable set V is that (i)there be a directed path from X to Y such that each link in this path is a direct causal relationship (…) and that (ii) there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value.

Intervention variable (Woodward 2003, p.98) has to meet several conditions:

I1. I causes X.

I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are sucha that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.

I3. Any direct path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are build into the I-X-Y connection itself, that is, except for (a) any causes of Y that are effects of X (i.e. variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.

I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

Interventionists apply their approach to causality, to the free will problem (List and Menzies, 2017), and argue that an event can be considered a cause in the interventionist framework if it satisfies, not one but two counterfactual conditions: (1) the positive: If C were to occur, then E would occur; and (2) the negative: If C were not to occur, then E would not occur. It should be noted that usage of those two conditionals in connection to mental causation is relatively broadly spread throughout interventionism, and not limited to List and Menzies approach.

According to the authors only the mental can be considered to satisfy both conditionals. The conclusion stems from the multiple realizability of the mental, where a change in mental always entails a change in the physical construction of the individual, while the converse does not hold, since multiple physical states can support the same mental state (multiple realization was discussed more in depth in the chapter on mental causation). List and Menzies argue that, although the brain state is causally sufficient for the overt action, it is the conscious decision and not the brain state that is the difference making cause of the action.

Parallels can be noticed between this approach to the causal exclusion of the mental and Schlosser's (2014) reasoning, when arguing that Libet's experiments establish a temporal order rather than causal links between the physical and the mental, since his requirement for causality is the possible intervention. The questions that remain, at least explicitly, unanswered are (1) if there are two types of causality at play here (the mental being a difference-making and the physical being some form of productive cause), and (2) if they would claim that there is some overdetermination at play here.

Interventionist's argument can be illustrated with a simple example (here is a modified version of the one discussed most often, presented also by Raataikainen, 2010): let's take Sally, a meat lover, who has been dreaming all day about a steak. She knows that in her fridge she has a beautiful cut of meat waiting for her that she bought the day before to eat as a pleasant Monday dinner. We can say then, Sally has a belief that there is a steak in the fridge. When she approaches to the fridge with the intention to fry the stake, is she guided by her belief that there is a stake in the fridge or some physical brain state? Interventionists firmly argue that it is the former rather than the latter, since there can be a number of physical brain states that Sally has, all of them being a supervenient basis for the belief that there is a steak in the fridge. Conversely, they argue, there is only one mental state that leads Sally to that fridge, the belief that it is where the meat she purchased is. If that belief changes, e.g. Sally starts believing that the meat is in a cupboard rather than the fridge, because she left it there to warm up, the action resulting from such mental state will change. Instead of going to the fridge Sally will turn to the cupboard instead.

Putting these intuitions in a more formal way, only manipulating the mental event seems to change what Sally will be doing: going to the fridge or to the cupboard. While, when we manipulate the physical brain activation, a supervenient basis for the mental event, we might still end up with Sally having a belief that there is a steak in the fridge, and ultimately leading her there instead of the cupboard. Putting this in the conditionals proposed by List and Menzies, for the mental event it holds that:

1. If Sally is in the mental state of believing that the stake is in the fridge, she will go to the fridge (the positive conditional)
2. If Sally was not in the mental state of believing that the stake is in the fridge she would not go to the fridge. Or if Sally was in a mental state of believing that the stake is in the cupboard, she would not go to the fridge [but rather to the cupboard] (the negative conditional).

On the other hand, for the physical event, the second conditional does not hold, as interventionists claim. If Sally were in a different physical state, that state could still support Sally's mental state of believing that the stake is in the fridge, which would result in Sally going to the fridge anyway.

The solution regarding the mental does not seem particularly controversial, in the end, as Woodward said, we often change other people's attitudes and through that affect their behavior. And it seems plausible to say that if Sally would believe that the stake was in the cupboard, instead of the fridge, we would expect that, if she wants to cook dinner, she would go to said cupboard rather than the fridge. Therefore, on the surface, it seems that in fact, when causality is seen as making a difference, what makes a difference to Sally's behavior is her belief about the location of the stake rather than any neural state she might have. However, it can be questioned if the proposed solution does not break the causal closure principle, as the physical cause can be seen as not sufficient. Moreover a similar version of this solution was explored by causal compatibilists, which I will discuss with more detail below.

We could interpret the interventionist solution as holding that (1) the mental is the difference making cause, and (2) that the physical is the sufficient cause, which would preserve the causal closure. This presents us with two possibilities (Raatikainen, 2010):

1. We are making claims about the causal relevance of one variable to another variable, e.g. the causal relevance of the mental state, believing that the stake is in the fridge, to the effect, going to the fridge to get said stake.

2. We are making claims about the variable's specific value, rather than a different value, as being a cause of a specific value of another variable, again rather than a different one.

In the first case, for a variable to be causally relevant to another variable in the interventionist framework, it is sufficient that changing said variable by some intervention will lead to a change in another variable, usually the effect. This is a very weak requirement, and in fact, it is possible that such an understanding of causal relevance is applicable to both the mental as well as the physical e.g. a variable can range over several possibly exclusive brain states, including the one which is the supervenient basis for Sally's belief that there is a steak in the fridge. Then some interventions on the physical can also lead to the changes in the effect. But in such a scenario the two: the mental and the physical, are not in competition in the same sense they are in the argument for the causal exclusion of the mental, as it is completely plausible to have several variables being causally relevant to the same effect (e.g. when a fire breaks in your house, as in Mackie's example, not only the spark, but also at least the presence of oxygen is needed for the fire to actually start).

This solution does not seem to satisfy the interventionists, as Woodward (2003, p.66) and Raatikainen (2010) after him argue, that this is not very informative. Instead of discovering, or confirming the existence of some manipulation on one variable, which then changes the outcome, they would like to obtain more detailed information on which interventions exactly will change the outcome. Most interventionists would therefore not be satisfied in simply concluding that the mental is somehow relevant to the overt action, as this is in fact not very informative, but rather would like to explain how the mental *rather than* the physical is the cause of an action.

This leads us to another inconsistency in the interventionist positions, this time regarding the notion of the cause. From one angle, it seems that the mental and the physical are seen as two competing causes, at least on List and Menzies account. However, Woodward himself states:

*It [*the paper*] defends the view that both supervening properties and the properties on which they supervene can be causes of the same effect. For example, I conclude that when mental property M_1 supervenes on physical property P_1 and mental property M_2 supervenes* *on P_2, then under the right conditions both M_1 and P_1 will be causes of mental property P_2 (and of M_2). It is a further question whether, in such circumstances, there nonetheless are reasons for distinguishing between the causal role of M_1 and of P_1 with respect to M_2.* (Woodward, 2015, p.303)

This would suggest that the two variables: mental and physical, are not taken to compete in the standard sense of the word, but rather they are both seen as a cause of an effect. This would bring Woodward's solution closer to causal compatibilists, who reject the premise of overdetermination.

Baumgartner (2009) argues that this, although according to him only theoretical, rejection of systematic overdetermination, is one of the things that are attractive for nonreductive physicalists. He substantiates his claim by referring to a case of a network of neurons where:

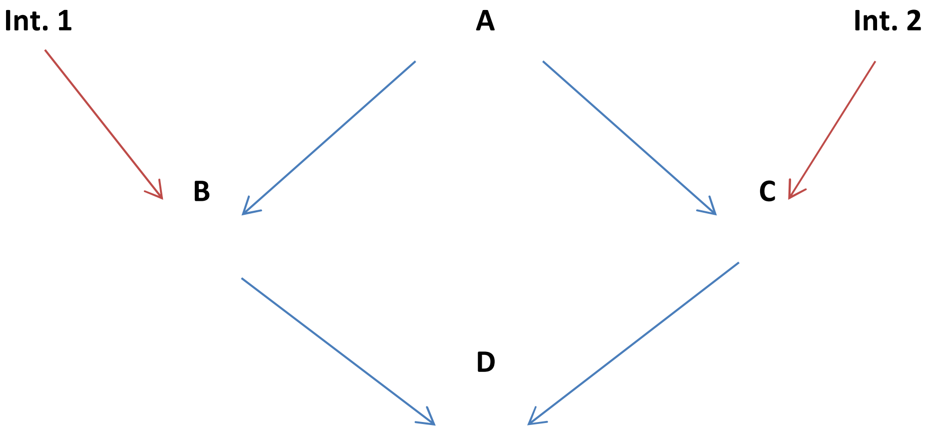


Figure 24. Neuron firing structure with the following properties: If neuron A fires both B and C fire as well. A firing of B and C, in turn, each is sufficient for neuron D to fire. Int. 1 and Int. 2 are two intervention variables, intervening on variable B and C respectively (Baumgartner, 2009, p.8[19])

In the structure represented above both interventionist conditions: fixability and manipulability, can be satisfied by intervening on either B or C while holding the other one fixed, and, in either case, we can assume that D will vary (since they both can be the causes

[19] An online version of the paper was used, from the author's website, therefore the page number.

of D). Then according to the interventionist framework, both variables can be seen as causes of D, and their firing systematically overdetermine D. Therefore systematic overdetermination, at least in some cases, could be allowed on the interventionist account of causality. Although the example above is illustrated using a neuron, the solution might not be applicable, as it is not clear if an overdetermination of mental of the physical, or conversely, would even constitute a cause of systematic overdetermination.

Does the solution proposed by interventionists actually shows how "disarm" the most potent argument against the causal efficacy of the mental? Some, most famously Baumgartner (2009, 2010), argue that because of the supervenience relationship that connects the physical and the mental together, the proposed solution not only does not work, but the required manipulations are impossible. And as we have seen in the previous chapter, for Baumgartner that would potentially mean falsification of the hypothesis of causality of the mental (and, surprisingly, of the physical as we will see further), or, at least, makes it impossible to establish if the connection between the two exists.

Baumgartner believes that, if we consider mental and physical to be two competing causes, as the consensus between interventionists seems to be, we need to be able to manipulate them independently, or manipulate one, while at the same time holding the other one fixed. This condition however cannot be met, due to the relationship holding between them (supervenience). As discussed in various places in this dissertation, the mental, on nonreductive physicalism, is not caused by the physical (nor causes the physical) but, in a sense, appears along with the physical (in Kim's diagram). This is forced by the core assumption of supervenience. The two also co-change, since two individuals identical in all physical aspects also have to be identical in all mental ones. On the other hand interventionism clearly defines direct and contributing causes by making such interventions, that change different variables independently, the defining feature of causality. For clarity the definition of Woodward states:

(M) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set V is that there be ***a possible intervention on X*** *that* ***will change Y*** *or the probability distribution of Y when* ***one holds fixed at some value all other variables Z_i*** *in V.* (Woodward, 2003, p.59; emphasis added by the author).

As I discussed previously, this definition is ambiguous, but we can assume that, from other arguments and general discussions, both the manipulation as well as holding other variables in the same set fixed was a necessary condition to establish that a variable is a direct cause of another variable. Woodward himself has stated:

Putting this in a form of a slogan, we can say that manipulability accounts are committed to the following: No causal difference without a difference in manipulability relations, and no difference in manipulability relations without causal difference. (Woodward, 2003, p.61)

Therefore some form of manipulation is necessary, which entails both the variable to be manipulabe, and the possibility to fix other possible causes (and we do not conclusively know what other possible causes are).

Why does the solution not work according to Baumgartner? Let's go back to Kim's diagram:

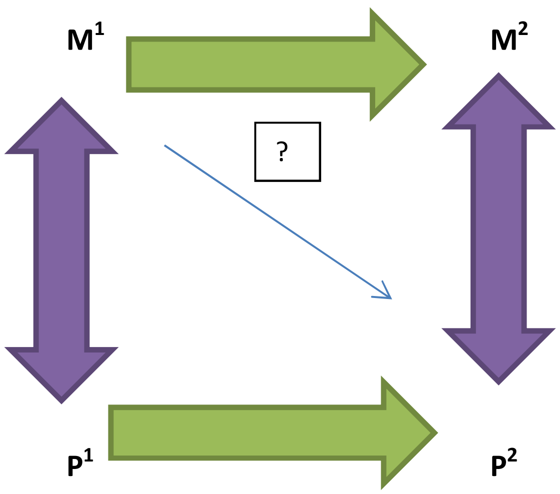


Figure 25. Kim's diagram. Mental to physical causation.

In order to show that M^1 is causal for P^2 we would have to be able to perform an intervention, in which it would be possible to (1) manipulate the M^1 independently from other variables in the set, and (2) keep the other variables in the set, especially other possible causes, at a fixed value. Other variables in this specific set are M^2 and P^1. Therefore there needs to be an intervention on M^1 with respect to P^2 that changes M^1 or its value independently from all the other variables, which could be causes of P^2 or lie on a direct path to P^2 without crossing M^1. Because of supervenience it is impossible to fulfill this condition.

As the two variables M^1 and P^1 are connected through supervenience, and moreover the mental supervenes on the physical, therefore any change in the mental necessitates a change in the physical, it is impossible to manipulate M^1 independently from P^1. Interestingly, because of the nature of supervenience, this is not the case for the physical. Since in the case of mental to physical relationship supervenience is considered non-symmetric, not every change in the physical has to necessarily be accompanied by a change in the mental, therefore this condition could theoretically be fulfilled, but there is no guarantee that it would be.

Not only can the mental not be manipulated independently from the other variables in the set, the other variables also cannot be fixed, again because of supervenience. Any manipulation done on the mental event entails a change in the physical realm, which is problematic also because the physical is the other potential cause of the P^2. If the two co-change when manipulated it is impossible to determine which of the changes in the value of the variables was the cause of the occurrence of P^2. Or, in other words, was it the change in the physical realm or in the mental realm? This is problematic not only because it shows that interventionism fails to show how mental can be causal, on this account it might be questionable if we can show that the physical is causal as well. This happens if we consider both manipulability **and** fixability necessary for a causal relationship. Although it is theoretically possible that manipulating the physical will not change the mental, since it is multiply realizable, the mental cannot be held fixed while we manipulate the physical. Even if its value does not change, it is coincidental, rather than a result of fixing its value.

Baumgartner’s challenge to some extent hinges on the necessity of the two conditions, the manipulability and fixability, although not completely. Even if those are not necessary for the existence of causal links, it is still difficult to see the way out of this conundrum, as, regardless if they are necessary, those conditions, which are in the heart of interventionism, cannot be met because of supervenience. In his earlier works Woodward (2003) seems convinced that both manipulability and fixability are necessary in order to determine the existence of causal links, however in his later articles (i.e. Woodward, 2008) argues:

I also assume that if a candidate causal claim is associated with interventions that are impossible for (or lack any clear sense because of) logical, conceptual or perhaps metaphysical reasons, then that causal claim is itself illegitimate or ill-defined.[…]Thus if we have two apparently competing claims, the first contending some mental state is causally inert and the other contending that it causes some outcome, it must be possible to specify some set of (coherent, well defined) interventions such that the two claims make competing predictions about what would happen under those interventions. (Woodward, 2008, p.224)

Unlike in his previous work, here Woodward states that if interventions, I assume such that you can both fix and manipulate independently, are impossible, the causal claims are not false but rather ill-defined. The new definition quoted above however, as Baumgartner emphasized, shifts the conditions cited in the definition from being the conditions of causality to being the conditions of meaningfulness of statements. Therefore the situations in which the independent manipulation is impossible a priori, for logical, mathematical or metaphysical reasons, the statements do not say anything about causality, but rather are ill-defined. This different version does not account for how the downward dependencies, necessary for mental causation, are possible. Moreover, it is unclear what we could say about mental causation, besides that most of the causal claims in such cases are ill defined.

However, not everything might be lost yet. The validity of the causal statements relies on the two conditions (fixability and manipulability), and both Baumgartner as well as Woodward seem to assume that there are only two, or three, possible interpretations of the claims involving two connected variables. In one of the previous chapters (on causality) I have identified at least five different, possible interpretations of such situations. Therefore there is still room for claims about causality, in spite of the inter-connectedness of the variables.

Does interventionism solve our problems?

Woodward (2015) defense of his position against Baumgartner's challenges was twofold: (1) he argued that the latter incorrectly uses Kim's diagram, treating the connection between the physical and the mental as causal, and (2) states that events connected through non-causal dependencies cannot be expected to be manipulated independently, with other variables being held fixed, since this is, a priori, excluded by the relationship that holds between them. Therefore, as Woodward argues:

At this point, we face a choice. On the one hand, we might (i) hold onto the idea that in order for M1 to have a causal effect on M2 (or P2), there must be a possible intervention that changes M1 while P1 is held fixed and under which there is an associated change in M2(P2) and infer from the fact that there is not such possible intervention that M1 does not cause M2(P2) […] Alternatively (ii)[…] we might conclude that the requirement that in order for M1 to have a causal effect on M2, there must be a possible intervention that changes M1 while P1 is held fixed and under which M2 changes, in inappropriate.(Woodward, 2015, p.333)

In other words, if it is impossible for metaphysical, logical or nomological reasons, to independently manipulate different variables, it should not be required, and scenarios in which it is, are ill defined. Because it is impossible to find scenarios of the first sort, Woodward argues that

By contrast, in the case in which one considers what would happen to M2 if one changes M1 while holding P1 fixed, the antecedent of the associated counterfactual is incapable of actually being realized- this is understood to be ruled out by inviolable constraints. Unlike the previous case, we do not argue for the causal inefficacy of M1 one the basis that the realization of the antecedent is possible, but that under such a realization the values of another variables would not change. Instead we argue for the inefficacy of M1 simply on the grounds that it is impossible simply on the grounds that it is impossible to realize the antecedent of the above counterfactual. That is, the version of the exclusion argument presented under consideration in effect invites us to conclude that M1 does not cause M2 (or anything else) from the impossibility of changing M1 while holding P1 fixed, while in the more ordinary case of control, we infer that X does not cause Y, by providing evidence that Y does not or would not change if some possible situation in which X changes while potential confounders Z are held fixed were to be actually realized. (Woodward, 2015, p. 336)

Woodward’s argument emphasizes that in two cases: (1) without non-causal dependencies between variables, and (2) with non-causal dependencies between variables, the assessment of the causal connection is based on very different reasons. In the first case, we conduct a manipulation, while holding other variables fixed, and conclude that the intervention did not change the effect, or its probability distribution, and therefore there is no causal link between the two: the variable and the effect. However, in the latter case, the interpretation very much hinges on what happens when the two conditions, manipulability and fixability, cannot be fulfilled. Woodward argues here, that the impossibility of these two conditions would lead to concluding that the mental cannot be a cause in this scenario. However, Baumgartner's argument might be interpreted to simply state that, under such conditions, and especially in connection to the assumption of empirical testability of the causal hypothesis, the causal link here cannot be established. In the latter case the tension between the two examples, that leads Woodward to argue that the non-causal dependencies have to be taken as a reason not to expect fixability and manipulability to be met, has very much decreased.

The second problem with Woodward's position, also mentioned by Baumgartner (e.g. 2013, 2016 for the basics) can be seen to lead to a paradoxical situation where what we choose to manipulate determines what we consider a cause. If the existence of the non-causal dependencies between the two variables excludes the independent manipulation, in principle we cannot, with certainty, assume it was the mental and not the physical, or conversely, that was the cause. Since they are both connected, the most we can say by using interventionism, is that there is a causal link from the cause, constructed from both the mental and the physical, to the effect. That can serve to determine if this, double-leveled cause, has links to the effect, while manipulating other, possibly also double-leveled causes. But it will not tell us, which one of the two inside the cause, contributed causally to the occurrence of the effect. If we will however insist that interventionism can lead to such conclusions, the latter will depend on which of the causes we choose to manipulate, automatically excluding the other cause. In the empirical applications, if one chooses to manipulate the physical, and through that reaches the conclusion that it was the physical, and not the mental that caused the change in the effect, such conclusion will be fully justified.

There are however ways to move forward from this deadlock, but they rely heavily on: (1) interpretation of the counterfactual possibilities, and (2) giving up the empirical testability of the results. The latter depend on which of the interpretations of the counterfactual possibilities we are willing to accept, as some of them simply exclude any empirical confirmation, and provide the basis for mental causation only theoretically. This would go in line with Baumgartner's (2017) suggestion, that the causal efficacy of the mental, is a metaphysical choice, rather than a modelling one. The first problem, and the different possibilities to interpret the counterfactual conditional were covered in the chapter on the modern approaches to causality. They include: (1) standard approach, (2) falsification, (3) the agnostic approach, (4) non-truth value approach, and (5) the counterpossible approach.

Interventionism and neuroscience.

One of the main claims of interventionism, that makes it so attractive to scientists, is the empirical applicability of its assumptions. However, as we have seen above, this might also be a vacuous claim, when applied to the mental to physical causation. The experiments are often conducted in a way that allows for manipulating different variables, which, in turn leads scientists to conclusions about the existence of causal links between them. As we have seen previously here, although neuroscientists rarely discuss their understanding of causality, because of the methods, and the articles that do discuss it directly, it is uncontroversial to assume that interventionism would be the best option for them. However, it is unclear if neuroscientists, also those working on free will, would commit to the understanding of causality interventionism offers, namely that causality is nothing over and above difference-making. The way brain connections are described, also in Libet's experiments, points to discovering, possibly through manipulation, mechanisms of how certain things happen in the brain. Therefore difference-making could be seen as a manner in which mechanistic connections are uncovered, and causality still might be defined as some form of production (as many of neuroscientific experiments in general aim to describe the neuronal exchange of electrical signals), either through quantity exchange, or through mechanisms. This could be further dictated by the, argued again by Baumgartner (2017), impossibility of generating any actual experiments that would prove interventionist claims, in an empirical manner, when applied to the mental causation problem. Baumgartner argues that, while possible in ideal circumstances, because of the supervenience relationship that connects the mental and the physical, no models can be generated that favor one over the other as a cause. Therefore a choice between the two causes is a reflection of one's metaphysical commitment rather than experimental evidence.

The main, implicit, assumption of Libet's experimental design is that the causes of our actions can be **either** mental **or** physical. Our movement of the finger is caused by either the readiness potential (or a different brain activation), or by the conscious decision. Most neuroscientists working on the topic, while finding reduction attractive (e.g. Haggard), still construct their experiments around the assumption that the mental is, somehow, different from the physical, that somehow the two are not reduced to one another. Through that they claim to have shown the falsity of dualism. This is clear. However under assumptions other than reduction or dualism, the way they see the ontological status of the mental vs. the physical is unclear. Moreover, they seem to ignore the fact that on at least some nonreductive accounts, the token mental is identical to the token physical, they only differ on the level of types. Since the experiments are conducted on tokens, there is nothing surprising about the existence of the brain activation, as was argued throughout, and dualism is unnecessary.

As was discussed throughout this dissertation, among the many different candidates for the cause that would give rise to an action in 'the right way', action that could be then interpreted as free, almost none are considered independent from the physical domain. Therefore neither the fact there is a physical brain activation prior to either the movement, or the conscious decision (as investigated in Libet's experiments), nor the possibility that dualism would be disproved by the results of the experiments, is a serious threat to free will. However, as was discussed in the previous chapters, the nature of the relationship between them should no longer be ignored, especially if we agree that the best explanation could be offered by nonreductive physicalism. Yet, up to date, possibly because of the implicit assumptions about the connections between the mental and the physical (the already mentioned assumption of dualism), no experiments are conducted on the relation between the mental and the physical, nor the consequences of this to the free will debate.

In the specific case of Libet-style experiments, it is hard to say if the readiness potential could be the supervenience basis of the conscious decision. Firstly, while not strictly required, it is often seen that supervenience relationship requires some form of simultaneity between the appearance of the physical basis, and the mental, supervening, event. Secondly, it seems that supervenience is only discovered through comparing two distinct entities, which would have the same mental properties but different physical properties. Or, in other words, establishing supervenience seems to require change, or comparison between two entities that have the same mental characteristic, and, if they differ, this would be the basis to state that the mental supervenes on the physical. This, in case of Libet's experiment, is simply not investigated. Therefore, if that is in fact the case, the type of relationship between the two cannot be established. However, as other experiments showed, the readiness potential appears without conscious decisions (e.g. the discussed experiments that included hypnosis), or is taken to be 'brain noise'. Therefore there are reasons to think that, since in both cases the readiness potential appeared without the mental event, it is not the latter's supervenient basis. The mental could also be an emergent property of the physical, and there is some evidence from the research on consciousness that could support such hypothesis (e.g. Dehaene, 2014). Emergence is a term that describes how some entities (emergent entities) 'arise' from fundamental entities. The emergent entities are not reducible to their fundaments. Currently the term is gaining interest, and is applicable in various domains, including engineering. In Dehaene's reseach consciousness could be taken to be an emergent property of the synchronous firing of the different areas of the brain. The synchrony is the only bridge between those, not connected parts. This can be, roughly, illustrated by an person looking at a tree, where said tree can trigger memory, language area of the brain, visual cortex, etc. and for us to perceive *a tree* all of those areas should fire in synchrony. Emergence comes in two varieties, synchronic and diachronic, therefore could be applicable also to the Libet approach, where the conscious decision would emerge from the physical brain activation in a diachronic manner. However, no longer could we argue just as strictly that the readiness potential is the cause of the behavior.

Another interpretation is possible here. Since the mental conscious decision appears seemingly 'from nowhere', not connected to the physical, we could argue that the latter is a supervenient basis of a different mental event. The decision would then also supervene on a different physical event, not tracked by the investigators. But again, also this is just a speculation, because of the inherent dualistic assumption; the relation between the two is not accounted for. However, while rejection of this assumption is uncontroversial, doing so without establishing what kind of relationship actually holds between the two, seems uninformative. With a problem investigated as in Libet's approach, but without the assumption of dualistic nature of the connection, the experimental results tell us very little about free will, and do not show that it is the physical rather than the mental that causes the movement.

Could the interventionist solution be applied in this case? There are two paths for us to take here, firstly we can assume that the two: the readiness potential and the conscious decision, are two independent processes; then the experimental question changes from investigating two competing, but connected, events, to two disconnected processes. The two different, independent processes can be easily manipulated, with both the conditions, presumably, necessary for interventionist to hold fulfilled. But this approach does little to answer the mental causation question, or in fact the free will problem. Moreover, as some (e.g. Schlosser, 2014) have stated that Libet's experiments only establish time sequence of the various processes, a hypothesis that is just as applicable to the situation where we are dealing with two, independent, processes.

Secondly, we can assume that the readiness potential is some sort of preparatory process, as is argued by e.g. Mele (2009), who finds the readiness potential to be a part of the intention forming processes. In such case, the real competition would happen between the conscious decision and its physical basis (if we are assuming the falsity of dualism and reduction, there has to be one), and all the problems described by Baumgartner still apply.

How about other experimental approaches? Desmurget et al. (2009), through manipulating the parietal and premotor regions of the brain (therefore the physical rather than the mental) in patients undergoing a surgery, were able to obtain a variety of results. It was possible to trigger intentions to move the contralateral limb (in case of the stimulation of the right inferior parietal regions) and intentions to talk (the left inferior parietal region). In cases where the intensity of the stimulation was high, the participants believed that they actually moved although no muscle activity was registered, and stimulation of the premotor regions caused the patients to move their mouth and limbs, while denying that they have in fact moved. We could question if the study conducted by Desmurget really proves something about the causal source of our actions and about free will. Just like with Libet's experiments we can question if the intentions, as well as the movements, obtained in such experiments can be seen as the same type as those usually discussed in connection to the free will debate. The answer to this question might again boil down to what we consider the relationship between mind and body to be. If we assume dualism, than the fact the intention can be evoked after manipulating the physical, can lead to a conclusion that the mental is causally irrelevant. However if we agree that there is a relationship between mental and physical, those conclusions are much less telling. It boils down to one question, do we find it surprising that there is a physical brain activation, potentially related to the mental?

The studies however provide good reasons to question if the application of interventionism, while excluding the necessity to manipulate independently between the two potential causes, in the empirical setting leads to the conclusions envisioned by Woodward and others, in line with Baumgartner's (2017) objections. It seems rather that in those cases the fact that, for the purpose of manipulation, the physical and not the mental event was chosen would a priori exclude the possibility to prove that the mental is the cause of the action, even if that would potentially be the case. It is important to emphasize that this approach works only when we try to determine which one of the two, and not both are the potential cause, and then the answer might depend on which one we chose to manipulate: the mental or the physical.

This experimental approach makes the question of the connection between the physical and the mental even more pressing, since there is a number of possibilities why a specific intention appears along with the stimulation of a specific brain region, including: (1) causation: the physical brain activation causes the mental intention, (2) the mental event emerges from the physical brain activation, and (3) the brain activation is the supervenience basis of the intention, therefore they appear together, linked. Yet, again, without this clarification not much can be said conclusively about the consequences of these experiments to the free will debate. Moreover, as in Libet's experiments, interventionism faces the same problems. Other studies (Bode et al. 2011) argue that the outcome of a free decision can be predicted long before it becomes conscious, again questioning the causality of the conscious decision itself. All of this leads to the famous slogan: my brain made me do it, claiming that the mental is not causal and is caused by a specific brain activation, but they are again based on similar assumptions, and the connection between the mental and the physical remains mysterious. Moreover, it very much depends on how we conceive of prediction here. We grow up with our brain, we learn, and there are processes taking place constantly. It could be safe to assume that, while mental could still be causally efficacious, and we are free, we make decisions that are based on those previous experiences. If we can predict what will happen based on these, is this really threatening? On the other hand, if we think of causality as backed up by the laws of nature, in the humean sense, and conceive of predictability in terms of Laplace, is this the type of predictability neuroscientists mean here? It seems unlikely, especially since, as was shown in Dennett's reasoning in the first chapter, we are unable to take the position of the all-knowing entity. Therefore the predictabilities' threat to free will is only a threat in some understandings of the word and not in others. Moreover, it seems that folk intuitions, I will discuss further here, do not classify the predictability of our action, as threatening to our free action.

On the other hand we could think that the RP is a supervenient basis of another, potentially not tracked by the investigator, mental event. It is not controversial to suspect that at the same time various mental events are "present" with only few of them attended to with full attention. Consciousness does not seem to be an on/off type of phenomenon, but rather gradual (Miller, Schwartz 2014). Therefore, we might be barely conscious of some mental events, while others are very much in the foreground. Likewise, the conscious decision could supervene on an underlying physical event that is not tracked by the experimenter. It is also possible that RP in fact is a supervenient basis of the conscious decision, but the subjective perception of the conscious decision comes with a slight delay (however this option is questionable and we will not discuss it further here). There are reasons to think however that the RP might be independent from decision making processes (e.g. the conclusions of Schurger et al. 2012).

Short summary.

Is interventionism a good solution to the causal exclusion challenge, in its application to the mental-physical interaction? It seems that, while with a lot of potential, it is unlikely that just a change of how we conceive of causality is going to help us. The main problem, for interventionism specifically, seems to be (1) the relationship between the mental and the physical, if it is supervenience, and that seems likely, the two co-change, and (2) thinking of the two as competing causes (in some sense). If they co-change, we cannot, in not perfect circumstances of the empirical tests, manipulate them independently, which seems to make a conclusive answer to the question ‘which one is it’ impossible. We can accept one of the possibilities I have suggested in the chapter on causality, however, those only apply to theoretical considerations, and not to real experiments on subjects.

If we take a look on the experiments, it seems that, while manipulation is their preferred tool, neuroscientists would not likely commit to the claim that causality is nothing over and above difference-making. Moreover when we use the difference making approach, as presumably Desmurget would, we could conclude that it is without a doubt the physical, and not the mental, that causes the action. Moreover it suggests that the mental is itself caused by the physical, although describes as ‘giving rise’ to the mental.

The discussion on manipulation brings to light two more potential assumptions, and points of criticism against the experimental approach: (1) why do we look at the decision making processes not as processes but rather as events, and (2) what is our notion of predictability, since it is used against the possibility of free will.

Rejecting overdetermination.

Another way to avoid the Causal Exclusion problem is based on rejecting one of the premises of the argument (mental causation, irreducibility, causal closure, exclusion or no-overdetermination). Each of those leads to a different result: (1) epiphenomenalism, (2) reduction, (3) dualism, and (4) causal overdetermination respectively.

Because of the consequences of each of the positions, either effectively making free will impossible, or leading to more problems, the most promising seems the rejection of the exclusion principle, which limits the causes to one. Again, the rejection is taken not to imply the simultaneous rejection of the causal closure principle, as the mental is only defined functionally and not ontologically.

There are many ways to change the way we think about causes in connection to the issue of mental causation, and through that free will. If we agree that the Causal Exclusion is a problem for free will, as it surely is for mental causation, one of the ways to avoid it, is to reject one of the premises of the argument. As I mentioned previously, rejecting physical causal completeness leads to dualism, moreover it is almost unequivocally endorsed by all physicalists, and is said to be supported by, and appeal to, the scientific advancements (e.g. Papineau, 2001). The principle of mental causation is essential for free will (as it allows for mental events to be causes), and supports the experience we all share, of us being agents. Irreducibility of the mental to the physical is also well established, and based on two principles: (1) the multiple realizability of the mental (discussed in the chapter on mental causation), and (2) the impossibility of mapping the mental onto the physical in a manner that follows strict laws (discussed in the part on anomalous monism, and advocated by Davidson). Moreover the three abovementioned principles can be accepted together, without any issues, but adding the fourth one, the no overdetermination principle, leads to problems. Accepting no overdetermination principle along with irreducibility and mental causation principles, leads to falsification of the physical causal completeness, which, while sometimes done (e.g. Lowe, 1993), is undesirable, given the stability and overall endorsement of the latter principle.

Therefore, one of the promising possibilities is rejecting the no-overdetermination principle, one explored by causal compatibilists, which leads to embracing some form of causal overdetermination. Again here, as with interventionism, two interpretations are possible: (1) there are two, in some sense, independent causes, the mental and the physical, and the mental does not overdetermine the physical, but is the cause of the outcome, and (2) the mental and the physical, while connected, cause something together, but they causal contributions differ, and, in some cases or under some understanding of causation, the dynamics between the mental and the outcome is more telling than the one between the physical and the outcome. The former is how most of the discussions are framed, while the latter will investigated in the second part of these two conclusive chapters.

A 'surface answer' (Bernstein, 2016) to what causal overdetermination is, treats two possible causes c1 and c2 as overdetermining the effect, if they are individually sufficient to cause the effect. Importantly, they are not only necessary, as would be on Mackie's account, rather they are both independently sufficient. According to Bernstein this encompasses many sorts of cases including, among others:

1. Two rocks overdetermination- the classical example, where two rocks are through at the window with the same force, at the same time and speed, hitting the window at the same moment. Each of the rocks is individually sufficient in breaking the window.
2. Mental/physical overdetermination
3. Mereological overdetermination, where the rock, thrown at the window is composed by parts. Both the rock, and its parts are sufficient to break the window.
4. Constitutive Overdetermination: where the rock is seen as composed from a set of minerals, and their sum constitutes but it not identical to the rock. Again, similarly to the mereological overdetermination, both the rock, and the minerals are sufficient to break the window.
5. Quantitative overdetermination- where the rock, braking the window, is (much) bigger than the quantity required to break the window.

With seemingly this many different possible types, we could question if there are good reasons to accept it in the first place, as a necessary assumption, fulfilling the conditions for free will.

On the other hand, there are five arguments in favor of causal exclusion principle (Moore, 2017): (1) it's an analytic truth (Kim, 2005), it is intuitively correct, (2) parsimony argument, or, in a nutshell, less is better- if one cause is sufficient, we shouldn't need any other causes, (3) coincidence argument- overdetermination is rare, while mental causation is ubiquitous, and such amount of overdetermination would be unacceptable (e.g. Kim, 1998), (4) additivity argument- if causation involves transfer, one sufficient cause, should transfer all the amount needed for the effect to occur, and (5) necessity argument- abandoning the causal exclusion leads to failure of one, the causal completeness of the physical, or both, the latter and mental causation, premises. The last point is most problematic, and it was used throughout this dissertation as the major challenge to causal efficacy of the mental. All of these are rejected by compatibilists, who usually argue that the mental does not overdeterminine the physical in the same sense, as in classic examples, two causes overdetermine one another. Therefore, the relationship between the mental and the physical, is different, in its causal understanding, than the two balls breaking the window at the same time. From the five different arguments against the overdetermination argument, if we still want to treat the two as independent causes, in my perspective is the fourth, since it connects closely to the threat of epiphenomenalism. The mental does not have to be reduced to the physical to be threatened, for free will, as I argued here already, it is sufficient it does not have causal powers. Which is the case in the additivity argument.

Compatibilists, proponents of the argument initiated by Horgan (1997) and Bennett (2003), argue that it is possible for one event to have both sufficient physical causes and distinct mental causes at the same time. In the case of free will, this would mean that it is perfectly plausible that a free action has both a sufficient physical cause, thereby sustaining the physical causal closure, and a distinct mental cause, the conscious decision at the same time. They usually posit a very close relationship between the two causes, mental and physical, and their effects (Yablo, 1992; Bennett, 2003; Shoemaker, 2007), where often the appearance of the physical cause necessitates the appearance of the mental. They claim that the relationship between the mental and the physical is not a case of systematic overdetermination, as the occurrence of the latter necessitates the occurrence of the former, based on their relationship. They usually endorse a counterfactual model of causation (e.g. Bennett, 2003, Arnadottir and Crane, 2013), where for p to be a cause of p* there needs to be a causal dependency between the two.

That causal dependency, stemming from the work of David Lewis (e.g. 1986), is based on counterfactual dependence in which two statements are true (Moore, 2017)

(I) *Had p occurred p* would have occurred*

(II) *Had p not occurred, p* would not have occurred*

As we have already seen in the previous chapters of this dissertation, the same counterfactual dependency is used by List and Menzies (2017), and they argue that only the mental can fulfil both of them, while the physical cannot. Thereby they conclude, on the difference-making approach to causality, from the two competing causes, it should be the mental and not the physical to be considered **the** cause. Quick recap of the arugment: only the mental can be considered to satisfy both conditionals, an argument based on the assumption of multiple realizability of the mental. In a nutshell, the mental is multiply realizable because the same mental event, for instance an intention to move, can supervene on several slightly different physical states. Since it is the mental that supervenes on the physical, a change in the physical does not necessarily entail a change in mental. List and Menzies argue that, although the brain state is causally sufficient for the overt action, it is the conscious decision and not the brain state that is the difference making cause of the action. They do not however agree that the mental overdetermines the physical, or at least it does not seem so, which would differ them from the compatibilists.

The proponents of causal overdetermination use the same counterfactual dependencies, without at the same time arguing for a necessity of a different concept of causality. In fact, which notion of causality they are using remains not completely clear, but some form of production, following Kim, or humean approach would be applicable. Overdetermination itself is an ambiguous notion, but the type in the mental to physical case is clearly distinguished from the genuine overdetermination used by Kim (as in the breaking of the window example).

The causal compatibilist position, from the get go, has a rough start. Superficially it seems obvious: the effect is overdetermined when it has more than one sufficient cause, and in the case of mental to physical causation, it is the case. The things are unfortunately not so easy, as Bennett (2003, p. 4[20]) argues that simply announcing that, of course, mental is causally efficacious, and can be considered a cause, would amount to stating ‘of course we have free will’, and ignoring all the arguments that show us it might not be so.

A number of issues can be identified here, firstly with the notion of sufficiency itself. When we look at Mackie’s proposal, which was covered in the previous chapters here, it might be impossible to ever identify one factor that would be completely sufficient, by itself, for the effect to occur.

[20] Also here I have used an article online (https://courses.cit.cornell.edu/kb383/exclprobrev.pdf), therefore the page numbers come from there.

Another question is, if overdetermination is what is happening in the mental to physical relationship at all, or if this is a completely different type of causal relationship. Most of the classic examples of overdetermination involve two independent causes, most used regard two balls hitting a window at exactly the same moment. In those cases it seems that each ball is, on its own sufficient, in breaking the glass, and therefore one overdetermines the other. In the case of the mental, if it in fact would be an example of overdetermination, it would be systemic, namely every mental event would possibly overdetemine every physical event, in cases that the two co-occur. The main goal of compatibilists seems to be to show that the mental does not count as a classic example of overdetermination, since the two cannot be properly seen as two independent causes, such as in the ball example. Then however, what is needed is a clear test, indicating which cases are examples of overdetermination and which one aren't.

Bennett (2003, p.8) argues that such a test can be presented in a counterfactual form, similar to the negative and positive conditionals used by the interventionists, and a Lewisian dependency, presented above. Bennett's test is also based on two conditionals, where effect (e) is overdetermined by two causes (c_1 and c_2) if and only if:

(O1) if c_1 had happened without c_2, e would still have [necessarily] *happened […]*

(O2) if c_2 had happened without c_1, e would still have [necessarily] *happened […]*

The test is designed to only provide necessary conditions for overdetermination. For Bennett, relying on counterfactual dependencies for the overdetermination test does not entail commitment to the Lewisian account of causality. She argues that Causal Exclusion is not linked to any specific understanding of causality, however, as was argued here previously, Kim was inclined to identify causality with a production account. Moreover the simple counterfactual dependence between the two causes is not sufficient to diffuse the overdetermination issue, since the worlds where the negation of c_1 necessarily entails the negation of c_2 are not relevant to either of the conditions presented above.

Another possibility is stating a closer connection between the two causes, where one cause guarantees that the other cause will also appear. In those cases it is metaphysically impossible for one to occur without the other one (which seems to be the case for the mental to physical relationship, at least on a token level), and in such cases the overdetermination counterfactuals are vacuous, and, as Bennett (2003) argues, the overdetermination is not guaranteed if both counterfactuals are vacuous. The latter however, given the multiple realization of the mental, is not true.

This point can be applied to the, already discussed, defence against causal exclusion used by List and Menzies, because only for the mental both of the conditionals (positive and negative) are true. Given Bennett's approach we could however argue that the falsified conditionals, on their account, are not falsified at all, but rather, shown to be vacuous. Falsifying them requires independent manipulation, which, as we have seen on the difference-making but also on any other account, is not possible because of the inner, non-causal dependencies between the mental and the physical.

We could question if this impacts the interventionist conclusion, and one way we could argue is that both conditionals have to be non-vacuous. Interestingly, Bennett herself, rejects the idea that seems to be at the core of List and Menzies reasoning: that the mental property could have supervened on a different physical property, or

If event m has occurred but event p not occurred, some other, closely related physical event p^* *would have occurred, and* p^* *would have been sufficient to bring about e. Similarly in the property case- if event c has property M, but had not had property P, then M would have been realized by some other physical property* P^**, and* P^* *would have been efficacious in c's bringing about e.* (Bennett, 2003, p.14)

The quote is relevant, because it seems that exactly the same property of the mental event, their multiple realization, is used by Menzies and Price to argue the mental, and not the physical, is the cause of an action. At the same time Bennett thinks that this sort of reasoning would result in the mental event being completely epiphenomenal. According to her, all it gives us is the mental being helpful in explaining why something occurred. Nonetheless the causal power stays on the physical level, since the closest possible world is not one where the physical is missing (since this would mean the world with free roaming souls), but rather it is replaced by another physical event.

Moreover, Bennett assumes that compatibilists should commit themselves to both counterfactuals being non-vacuous. Therefore the challenge, presented above (the vacuity of one of the counterfactuals), works better when applied to the argument by List and Menzies, but is not a solution for the overdetermination counterfactuals.

In the end Bennett's solution relies on how strictly we define causal sufficiency. According to her:

Suppose first that we adopt a rather strict notion, according to which the only thing that counts as causally sufficient for an effect is a whole big package, consisting of what we have intuitively thought of as a cause plus all necessary background conditions and causal intermediaries by means of which it brings about its effect. On that kind of notion, only rather complicated and partially extrinsic physical events or properties will ever causally compete with mental ones, and thus will ever be invoked in substitution instances of (O2). And such physical causes will indeed necessitate their mental competitors, and the relevant counterfactuals will come out vacuous. Now suppose that we instead adopt a rather more permissive notion of causal sufficiency- a notion on which an event or property can count as causally sufficient for an effect even though it requires some set of background conditions and intermediaries in order to bring effect about. On this kind of view about causal sufficiency, much more intuitive physical events and properties will be invoked in substitution instances of (O2), and those substitution instances will not in general be vacuous. However, they will typically be false, for precisely the reasons rehearsed above. Thus whetehr causal sufficiency is understood in the strict way or the permissive way, the compatibilist has an answer. (Bennett, 2003, p.25)

The first understanding of causal sufficiency seems to be close to, again, Mackie's INUS conditions, who assumes that there will be no one necessary and sufficient cause for events. Rather a causal field should be considered where a variety of necessary but not sufficient causes will contribute to the final outcome.

Some (e.g. Moore, 2012) argue causal overdetermination comes in two varieties: (1) independent overdetemination, which suggests that two causes, that are individually sufficient for one effect, bring that effect about, and (2) dependent overdetermination, where a physical cause is necessary and sufficient, but it necessitates a mental cause of the effect.

Following the first route leads us to a rejection of the causal exclusion principle, where we accept that every event has a sufficient physical cause, and a distinct mental cause, and we allow for events to simply have more than one sufficient cause. This is the solution chosen by, as we have seen above, Bennett (2003), but also argued for by Pereboom (2002). The latter is particularly interesting in the context of free will, since, as we have seen previously in the initial chapters of this dissertation, for other reasons, he argues against the existence of free will. While, if we take mental causation to be essential for it, his position on causal compatibilism, could be used to argue for it.

On the independent overdetermination solution mental causation is seen to operate somewhat similarly to the events usually quoted as examples of overdetermination, like the firing squad. On those both, the physical and the mental event, would be independently sufficient for the effect. Then both the occurrence of the physical, as well as the occurrence of mental (in a world with free roaming souls, independently), would bring about the same effect.

Moore (2012) argues that the independent overdetermination solution will not work, and one of the reasons for it is the same as the one used in justifying the existence of the principle in the first place, namely the amount of overdetermination that would have to exist as a result. As stated previously, mental causation is ubiquitous, and if mental overdetermines the physical, it would have to do so thousands of times a day, which seems to be against the odds.

Secondly, accepting independent overdetermination, might lead to the causal completeness of the physical being undermined. If the physical cause is sufficient, it seems no other cause is really necessary, however it seems permissible that, based just on this premise, two or more sufficient causes exist (depending on how we define sufficient). Yet, if the mental cause is sufficient, we could also conclude that the physical cause is not sufficient, which would violate the premise of causal closure directly. If we take an example of an action, e.g. a movement of a hand, the situation would look like this:

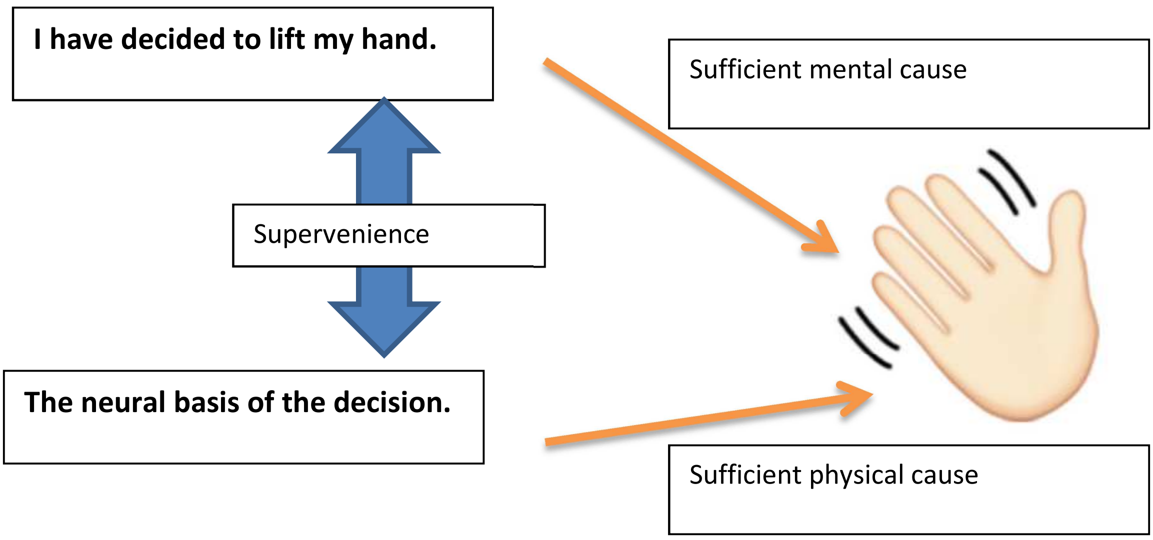


Figure 26. Independent causal overdetermination of an action.

As illustrated above, each of the possible causes would be taken to be sufficient on its own. Therefore the mental would be sufficient for the hand movement, seemingly excluding the physical, suggesting that in situations where no physical cause would be present, the mental on its own would be enough to bring the hand movement about. This again brings us in the midst of dualism, since it assumes that, on some level, there are situation where the decision to lift the hand, without the neuronal basis, would cause the movement. Or, that there are, relatively close, possible worlds where the souls roam free, which seems unlikely. It also permits a neural activation on its own to be sufficient for an action, something that I am arguing against, as it would further undermine the existence of free will, as it would mean that the same, exact action, could, in principle, be caused by just the physical, unconscious process.

The second type of causal compatibilist, similarly to the independent overdetermination, also endorses the idea that we should reject the no-overdetermination principle, but, in doing so, does not claim that the two causes are independent from one another. Often their arguments are based on emphasizing the differences between the mental/physical causes, and other, classical types of overdetermination, such as the case of throwing two rocks and breaking the window, or the shooting squad killing the prisoner. Barry Loewer (here I quote after Moore, 2012, p.324) argues:

In the two assassins case the two causes are metaphysically (and nomologically) independent. In the latter case M depends on N since it is metaphysically (or physically) entailed by it… The reason that we think that the first kind of overdetermination is generally not plausible is that when we posit two independent causes C1 and C2 for E we raise the question why both C1 and C2 and why the events that occurred led to the same effect. But in the case of [Nonreductive Physicalism] *we don't face these questions since N and M are not independent* (Loewer, 2002, p.658)

Therefore to clarify, the two assassins case could be illustrated as:

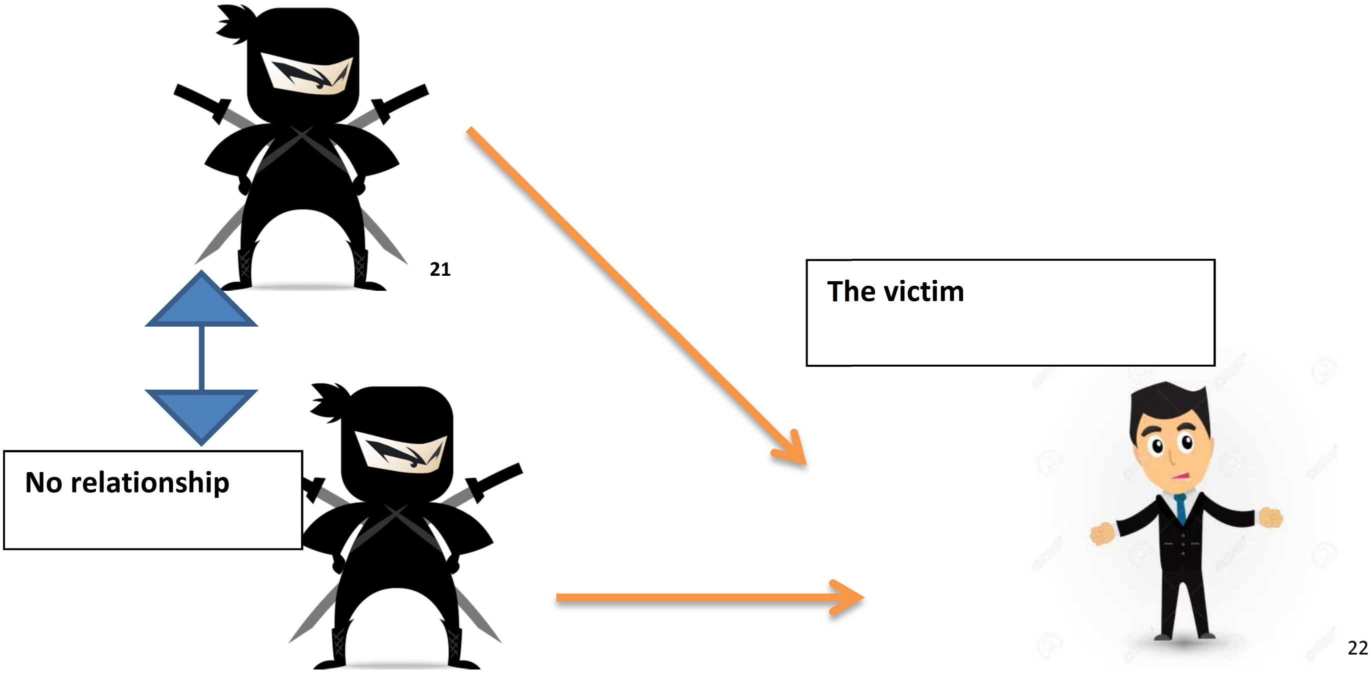


Figure 27. Independent causes, not connected via supervenience.

In the example illustrated above, two assassins shoot (or, in the illustration above, sword) the victim at the same time, both bullets/swords reach the heart at the same time and the heart of the victim stops. Each of the shots/stabs would have been sufficient to kill the victim. We can immediately see the differences between the two situations. In case of the mental and physical, where they are taken to be competing causes, they also do not

[21] I took the picture of the ninja from: http://tex.ninja/
[22] The picture of the man was taken from: https://www.123rf.com/photo_43777359_stock-vector-business-man-cartoon-vector.html

appear independently, unlike the two assassins who share no relation. Therefore only the latter can be really considered independent, while mental is always, necessarily, accompanied by the physical (supervenience). If we accept the strong version of supervenience (discussed in the chapter on mental causation), the conclusion follows that the occurrence of the physical basis necessitates the occurrence of the mental event (e.g. Kallestrup, 2006, Walter, 2008, Moore, 2012).

Dependent overdetermination is then defined as follows:

[…] some cause P is necessary and sufficient for effect e, but this cause P renders it necessary that event M occurs, while mental causation renders it necessary that event M causes effect e, so this cause M necessarily occurs, and necessarily causes effect e as well. (Moore, 2012, p.324)

This definition seems in line with Kim's position regarding mental to physical overdetermination, since he himself argued that the former overdetermining the latter cannot be seen as a classic case of overdetermination (e.g. Kim, 2005). On the dependent overdetermination account the physical causal closure principle, at least at a first glance, seems to be intact, since the physical cause is necessary and sufficient, while its appearance will also necessitate the appearance of the mental event, which then will cause the effect. This approach not only preserves the causal closure, but also avoids the seeming randomness of the overdetermination, when we treat the physical and the mental as independent causes. In the latter there is no explanation of why the two causes appear together, while on the dependent overdetermination account, the co-occurrence is explained by the necessitation of the mental by the appearance of the physical. Moreover, the mental necessarily causes the effect, because of the mental causation premise.

Promising as it seems, the account also has its flaws. First set of problems stems from the causal powers. As Bernstein (2016) argues, it is unclear which are the causal powers of the mental and those of the physical. In case of independent causes, at least in principle, we can show what part of the effect was caused by which cause, and here we do not have this type of the picture available to us in case of mental to physical causation. To illustrate this, the neuroscientific experiments on human volition are useful. Even if we assume that the conscious decision, as in Libet's experiment, supervenes on the readiness potential, since we are not able to distinguish between them, we could still argue that the movement of the finger was due to the physical only, with mental being epiphenomenal. Again, to hold that the mental caused the effect, we need to argue that it caused it qua mental, and not qua all the physical causal powers.

Second problem has to do with the dependence of the mental on the physical. Since the former always appears, and is necessitated by the latter, it is impossible to determine if the physical is individually sufficient for the effect, since the two do not occur independently, and are not treated as independent in any way. If we apply the counterfactual analysis onto this argument, the compatibilists can again either argue that one of the counterfactuals is false, or that it is vacuous. Yet, on this account it might turn out to be impossible to confirm the individual sufficiency of the physical cause. These problems suggest that, while this is not the end of compatibilism, more work needs to be done, or another approach developed.

Some lessons from the experimental philosophy

I have discussed the experimental approach to the question of the compatibility of free will with determinism in the first chapter of this dissertation. While most of the energy there is still spend on this issue, and some (e.g. Nahmias, et al. 2014a, Nahmias, 2014b, forthcoming) attempt to incorporate the presumable threats to free will from the conclusions of neuroscientific experiments. Interestingly in these experiments, the neuroscientists are taken to understand causality as law-governed co-occurrences, much in Hume's style.

One of the possible consequences of the neuroscientific conclusions, where the action is caused by the physical brain activation instead of the mental cause is the potential perfect predictability of the action, from just the physical recordings. This could potentially threaten free will in the same way as determinism does, as the predictability is the consequence of it. However, certain independence between the issues has to be drawn here. The problem of causality, or more specifically mental causation, as discussed here, is independent from the problem of determinism. Mental can be causally efficacious, but causing the behavior in a deterministic way. An action could be deterministic, and still not be predictable, if in order to predict it we need laplacean- demon style knowledge, which is inaccessible to us. Moreover, predictability can be a problem both for physicalists as well as dualists, where the behavior can be predicted either on the basis of brain scanning (in the case of physicalism), or mindreading (in the case of dualism). Nahmias investigates both of these scenarios, additionally adding, in one of the three experiments (Nahmias et al. 2014a) sentences suggesting that the mental can be reduced to the physical. Thereby he effectively presented the participants with three potential responses to the mind-brain relationship: mind-brain identity, nonreductive physicalism, and dualism. What should also be noted here, the causality that is taken to underlie the connection between the cause and the effect here, is the law governed correlation, mostly similar to the humean understanding of what causal relationship constitutes of.

The results of the experiments showed, with statistical relevance, that the participants rejected the hypothesis, that the predictability threatens the free will. In all cases, also in the identity-scenario, the fact that a technology can exist to predict the action of an agent, said action was nonetheless deemed to be free. This changes when, instead of prediction, manipulation was used. If the brain activation was, or could have, been manipulated, the attribution of free will was different, with most people thinking that, in those cases, the agent was not free.

The latter conclusion can lead to different possible interpretations. Nahmias argues that most people are 'theory light' with regards to their theories on mind-brain relationship, or, in other words, they do not commit themselves to any specific theory of the mind-brain connection (e.g. dualism, reduction or nonreductive physicalism), nor to a specific manner in which the mental states connect to the physical states (e.g. Nahmias, 2014b).

And, while it might be in fact so, there is another interpretation, one that I will explore in the coming chapter, namely that people do not usually consider that there are two causes at play here. Rather they consider it to be one, complex cause, composed from two properties: the physical and the mental.

Conclusions.

After establishing that mental causation is one of the necessary conditions for the existence of free will, and examining its biggest challenge, the causal exclusion, in this chapter I have analysed the, currently most popular, solutions, that allow for, in a way having your cake and eating it too: (1) thinking of causality in terms of difference making, allows to argue that the mental is causal, in the interventionist sense, and (2) the physical is causally closed, but the mental, defined functionally, overdetermines the physical. Both of those solutions, as I argued here, have major shortcomings.

Interventionism can be in fact considered a type of causal compatibilistm, at least in the List and Menzies version, where the two counterfactual conditions, and the rejection of no overdetermination principle are discussed together with the different approach to causality. However, two major problems appear here: (1) it is unclear how, on their account the physical causal closure holds, as their solution is closed to the independent overdetermination than to its dependent variety, which seems to suffer less problems, and (2) if Baumgartner's objections are valid, and I think they are, the impossibility of the independent manipulation and fixability makes experimental investigation, the core of interventionism, impossible. Because of the latter problem, mental causation becomes a metaphysical commitment only.

Causal compatibilism, coming in two varieties: independent and dependent, uses similar techniques, also advocating the overdetermination of the physical by the mental. The independent solution also seems to violate the causal closure principle, while the dependent solution does not seem to determine if the mental to have any causal power.

If, as the problem of free will seems to require, we need the mental to be causal qua mental, none of the proposed solutions will do. However, the results obtained by experimental philosophers seem to suggest another solution: that what we, intuitively, consider a cause, is different than what is argued for in both neuroscience and philosophy. Throughout this dissertation we have seen the way mental and the physical are opposed, competing to be the cause of the hand movement, or a more complex behavior. And yet, we are eager to accept the flat ontology of nonreductive physicalism, where the mental and the physical event are token identical. Therefore it might seem, and I think it does, that the main issue lies not in the way we define causality, but rather in the way we approach the notion of the cause.

Chapter 10. The new, unexplored ways to move forward.

After establishing, throughout this dissertation that:

(1) Causality, determinism and predictability, while interconnected, can cause different problems :

  a. Determinism, understood as causal determinism, shows that our future is causally closed, and there is only one path for us to take.
  b. Predictability can be understood in various ways, and holding with various strengths. If we understand predictability in terms of a laplacean demon, who, with the knowledge of all the antecedent states of the universe, and the laws of the universe, can in principle predict the whole future. Dennett argues that this is not achievable for us. On the other hand we can also understand predictability as we do in sentences 'I knew she will choose to study philosophy, I know her so well', or think that the brain is, as are the mental events, formed on the basis of previous experiences and interactions, and the decision making is a process that at some point becomes predictable. The two latter understandings are much less threatening to the notion of free will, than the first one.
  c. Causality- the action, to be intentional, has to be caused in the right way.

Here I argued that the three can be discussed independently, and, can create separate problems. This led me to analyzing what the 'right causes' for the intentional action are. I have identified several possibilities here: (1) intention, (2) conscious decision, and (3) reason, and I have investigated different types of those, and connections with other concepts, such as, e.g. goal. The characteristics that seem to connect them all was: (1) their mental character, and (2) some form of connection between them and the physical realm. The last excludes dualism, which is assumed in the neuroscientific experiments. I have also suggested that, from other possibilities of describing connection between the mental and the physical, also reduction is unlikely. This leaves us with some form of nonreductive physicalism, which can accommodate both the intuitions from the philosophy of action, as well as the information from the experimental realm.

I have then argued that, given the overwhelming evidence, a necessary, but not a sufficient, condition for the possibility of free will is the causal efficacy of the mental, which, in turn, is threatened by the causal exclusion.

In the last chapter, the first of the concluding ones, I have presented, and argued against the two major approaches to solve, or avoid, the causal exclusion of the mental, that at the same time leave room for the possibility of free will. I have shown that, in the specific case here, none of the positions allows us to show that, either, mental causation is possible, or that the type of mental causation we need for the possibility of free will to be preserved, is possible.

In this chapter I want to explore the possibilities that could allow us to move on. The problem has been a deadlock on the talks on causal efficacy of the mental, and a solution is needed. The ideas presented here are not completely worked out, and are here to create a fundament for a position to come.

How many causes are there then?

In the final sections of the previous chapter I have discussed the new experimental approaches to the problem of free will in philosophy, mainly in the work of Nahmias. There he, in two different studies, discovered that: (1) predictability is not a problem in ascription of free will, even though the participants heard that there could exist a machine that predicts the behavior of an individual, said individual would nonetheless be free, (2) the fact that brain activations exist, and takes part in the generation of a behavior, is not a problem for the existence, and ascription of free will, and (3) only manipulation of the brain activation, based on the prediction, is seen as problematic for the ascription of free will, therefore if the brain activation was manipulated, the individual no longer had free will, according to folk views. I think this happens for a very specific reason, namely, that the individuals do not believe that there are two different causes: the physical and the mental, at play here. By that I do not mean or suggest that the folk intuitions are reductive, quite the contrary, they do believe that both the mental and the physical exists, without necessarily having a full theory on how that happens, which Nahmias (2014b.) calls being theory light. However, I think that the folk intuitions might lead to a different understanding of what a cause is in the causal links connecting behavior and its cause. Intuitions that I think are present also in neuroscientific as well as philosophical discussions, when not framed in the context of mind-brain relationship specifically.

When intentional action is discussed, it seems to be **identified** through the existence of the intention. Therefore, from the purely physical perspective, the intentional arm movement and the automatic arm movement can be identical in all aspects, besides the existence of the intention. Yet, when we move to the discussions on mental causation and the mind-brain relationship, almost equivocally, everybody treats them as competing causes. A claim that can also be seen as strange, or at least those are my intuitions, as the two are not defined in the same manner: the mental is defined functionally, and the physical ontologically. Therefore ontology is flat, and, at least for me, this suggests a residual dualism in all positions that are trying to still define the two as competing causes.

The connection between the mental and the physical is at the core of nonreductive physicalism, however the nature, as well as the strength of this connection varies. As we have seen throughout this dissertation such relationship, while usually defined as supervenience, has been also described as dependence or realization. Supervenience is probably best defined and most commonly used, with the latter two more ambiguous (e.g. with the notion of realization, we have seen that, while there is a significant body of literature concerning multiple realization, the concept of realization seems to be much less investigated). Also in the discussions on causality, the exact relation connecting the mental and the physical events is unclear, with some describing the former as being constituted by the latter (e.g. Pereboom, 2002), or realized by the latter (e.g. Shoemaker, 2001, 2003). Moreover the exact extent of their causal powers is unclear, from sharing causal powers, the mental inheriting causal powers, etc.

Three major assumptions, that rarely seem discussed are important here: (1) that the mental is somehow the same as physical, or at least that it's defined in the same way when discussing the mental and the physical causes, (2) that mental and physical causation have the same fundamental characters, and (3) that the mental and the physical are two competing causes. However, the discussions on intentional action and the results found in the studies in experimental philosophy, lead me to believe that, in the discussions on causal efficacy of the mental, and especially in its application to free will, we are wrong about all three of them. Therefore, in order to find a conclusive answer to the issues discussed in this dissertation, we have to change how we think about the mental, the physical, the relationship between them, but, most pressingly, about what we consider a cause.

The first point is intuitively problematic for me. The physical causes are, or at least seem to me, often defined through their spatial-temporal properties. Defining what the cause is, from the point of view of empirically testable science, seems to however be unavailable for their mental counterparts. The example of free will touches the core of the problem: the implicit assumption of neuroscientists, as I argue here, is both that mental has to be efficacious for us to have free will, and that it is not connected to the realm of physical, and that this connection is to some extent irrelevant. This, at least to me, entails a certain level of commitment to treating the mental as, at least, definable in the same manner as the physical: ontologically. However, if we are not committed to dualism, but rather are in favor of nonreductive physicalism, and I have shown that this is our best options if we want to defend free will, incorporate neuroscientific results, and believe that mental needs to be causally efficacious to do so, only the physical events are defined ontologically, while mental functionally. I believe that the treatment of the two of them, as definable in the same terms, leads to two other issues I discuss here.

The second point, has been discussed previously in the context of causal exclusion (Menzies, 2003), where he argues that the concept of causation is treated as the same as applied to the mental and to the physical, which can yield a conclusion that the two are relata of causation (after Crane, 1995), and not causation itself, which is how they are usually interpreted. Rejecting this assumption is one possible route to reject the consequences of the causal exclusion argument. I disagree that we should reject the idea that the two, mental and physical, are relata of causation, but I believe that, because we identify the physical and the mental, based on different analysis (the former ontologically and the latter functionally), this should impact the way we see causation, when we discuss either one of them.

Mental causation is a term used when **at least one** of the relata of causation is mental, while physical causation means that **at least one** of the relata of causation is physical (Crane, 1995[23]), but is the relation itself different? The causal exclusion argument, or at least the way it is formulated by Kim, treats the causation as production, but very few other theories, as we have seen above requires that the type of relationship that connects the two has to be physical. This is a step in the right direction, but as we have seen the previous solutions, such as, most notably, interventionism, fail to provide a good solution, on which the causal closure and irreducibility of the mental would remain, while explaining how mental could be causal under these conditions.

The third issue here is, in my view, a major factor in the failure in the various solutions of the causal exclusion argument. Most of them, including interventionism and causal compatibilism, are based on counterfactual analysis, which calls for, at least conceptual, independence of the two potential causes, as it assumes a failure of one of the counterfactuals. Even those, who argue that the connection between the two excludes such possibility, like Woodward, still the two causes are considered somehow independently. As we have seen in the previous discussions Woodward still upholds that the relation between the mental and the effect might be more interesting than the one between the physical and the effect. This still seems to assume we are able to distinguish them as two, however interconnected causes. As we have seen in the case of dependent overdetermination, this solution also has its serious shortcomings.

I think that those positions are relying on a wrong idea what a cause is in the case of mental causation. It seems that the way we perceive a cause is, just like in the notion of causality, a matter of an agreement, however it seems rarely questioned, at least in the case of mental to physical relationship, if we are dealing with one cause with two properties, or relata, or with two causes. John Searle famously stated that our tradition is heavily influences by the dualism of mind and body, which makes it difficult to think of

[23] https://philpapers.org/archive/CRAMC-2.pdf

neurobiological processes as having conscious and conceptual properties (Searle, 2002a, 2002b), and he called this approach conceptual dualism (Searle, 2015). We have seen how the grip of such dualism in the neuroscientific experiments, however it seems that it has also influenced our idea of what a cause is.

An interesting duality can be observed in the philosophy of action discussions, where intentional action is taken to be one caused by a combination of the physical brain activation and an intention. Without the latter, the action would just not be considered intentional, an assumption repeated also in the neuroscientific approach to intention. It is my opinion the interpretation of this is twofold: (1) there are two causes, that, together, cause the action, or (2) there is only one cause, with two types of properties, physical and mental. On all accounts I have analyzed, the first one is assumed, and what the authors are looking for, is an answer on how the mental cause can be the effective cause, while being linked to the physical cause. Here I want to explore the other possibility, which, to some extent, resembles the causal compatibilist approach, in recognizing that the link between the two: the mental and the physical, determines the type of account we can have (the physical cause is there, without a doubt, the problem is how to accommodate that there is also the mental cause).

My position however differs on what can be called a cause here. Unlike my predecessors, I want to argue that only one cause exists, consisting of both mental and physical properties. Both of those properties contribute, but a cause, in the classical sense, is only one. This allows to accommodate the intuitions seemingly found both in discussions on intentional action in general, as well as the intuitions of the folk.

The accounts which call for solutions, where the mental competes with the physical, as we have seen throughout here, consistently fail to provide a solution to the Causal Exclusion problem, and encounter many, seemingly independent from the latter, issues. My suggestion avoids that competition altogether, as the two do not compete with one another, but are just ‘parts’ of one cause. This also seems a more natural way of capturing the difference in defining of the two: the mental, functionally, and the physical, ontologically. On all other examples of non-related causes, both of them are defined in ontological terms. Therefore my account would only apply to situations where the two properties, usually treated as two causes, are related to one another by links like supervenience. Since I am only attempting to sketch a basis of an account here, as it requires much more work, I will limit myself to the mental/physical case, in its application to free will specifically. At the core of my intuitions are Mackie's INUS conditions, but instead of rejecting the idea of necessary and sufficient cause in favor of the causal field, I want to argue that the necessary and sufficient cause of a voluntary action, has necessarily two properties: the mental intention, and its physical supervenient basis. I hypothesize that another reason for this approach, instead of the competing causes one: when we accept the mental nonreducibility principle, we are not doing so, while simultaneously claiming that the mental is independent from the physical, or can exist in any way outside the physical realm. Therefore we do not treat it as independent, yet when we focus on causality, the same cannot be said, as we seem committed to, at least implicitly, assume that the two are independent as causes.

My solution also avoids another major problem for free will, namely epiphenomenalism. On my account, since the cause is only one, mental is causally efficacious inside the cause, but cannot be taken to compete. The causal powers are on the side of the cause, not on the side of the properties.

This way of thinking about a mental and physical seems to resemble the way we think about information, where the semantic part of it, which supervenes on the physical part, is usually conceived of one cause, not two competing causes, at least as far as I know. Here by the semantic part I mean the meaning, or the aboutness of the mental, in the most general sense. Moreover, currently under development are new approaches to causality, involving information transfer, which could, eventually have the potential to account for the complex causes, I have in mind here. Another possibility, that would allow for such types of causes, might be some form of diachronic emergence, where the mental is an emergent property, appearing from the physical at some point, and influencing a change of laws, where the novel phenomena appears throughout time (e.g. Humphreys, 2008). Therefore, while the mental events are 'caused' by the physical events, the former's powers and behavior cannot be fully explained by the latter. The emergent nature of, e.g. consciousness, was suggested also by Searle who stated:

Consciousness is caused by a lower-level neuronal processes in the brain and is itself a feature of the brain. Because it is a feature that emerges from certain neuronal activities, we can think of it as an 'emergent property' of the brain. An emergent property of a system is one that is causally explained by the behavior of the elements of the system; but it is not a property of any individual elements and it cannot be explained simply as a summation of the properties of those elements. (Searle, 1997, pp.17-18, after Vicari, 2008, p.50)

In the following closing remarks I will only explore the existing positions of causality understood as information transfer, for two specific reasons: (1) I do not feel well versed on emergence, and (2) I am aware that emergence also has its own problems. I therefore think that causality as information transfer might be more promising, and it may include emergence in the later stages of development of the theory.

How to account for complex causes.

As we have seen throughout existing accounts of causation suffer a number of problems, including absences, not properly dealt with on any of them, and understanding causation as information transfer offers the hope to solve those problems. Secondly there are reasons to think that the latter understanding might have the advantage of fitting into the existing accounts of how the brain works.

The disadvantage of the current accounts of causation as information transfer, are investigating exclusively cases of physical causation, therefore it would still have to be worked out, how does the semantical content, and ultimately the mental, be accounted for here. However, the notions presented here can form a fundament, used to build the position further.

The way we conceive of information, we usually think of a piece of information, which has both the 'physical' and the 'abstract' properties to be causal, although not on all accounts of causation, as we will see further. However, the crucial part of this position is the concept of one cause, if we require that the mental competes with the physical, the position most likely suffers the same issues as the other ones: the mental is epiphenomenal if we don't find a solution to the Exclusion argument. The same problem does not apply for the modification, where we are dealing with a complex cause. In Woodward's words, there are still interesting things to say about the mental and the physical, and their mutual relationship, but they should be discussed in disconnect to causality.

Causality as information transfer.

Philosophy of information is a new and emerging domain, and while, exciting, it is also underdeveloped. As for causality understood as information transfer, we could even ask if another account of causality is needed, since so many can already be found. And, as we have seen previously two questions are relevant, in connection to causality, (1) what constitutes a causal link, and (2) how can we find out if the link we have observed is in fact causal. When it comes to the second question, the answers seem very much contingent on how we pose the question, with Illari and Russo arguing:

There is an interesting kind of question that scientist often ask when they are trying to find out whether C causes E. Physicists or electrical engineers might ask can energy be transmitted between C and E? Can radio waves? Can electrons? Software engineers might want to know whether bits can be transferred quickly enough for a program to run. Doctors might ask could this virus have been transmitted over such a distance so quickly? (Illari, Russo, 2014, p.135)

It seems that at the core of such questions is the notion of transmission, much like the idea that's at the core of Salmon and Dowe's solution, I have discussed previously. In these approaches, a causal link between the cause and the effect is presumed, and we can rule out which one of the possible causes, is the de facto cause, but showing that it has a link with the effect. Moreover, as was mentioned previously, also Jaegwon Kim understood causation as some form of production, and one of the ways to reject the Causal Exclusion argument was to change how we conceive of causation (e.g. interventionism, and causation as difference- making). What Russo and Illari seem to suggest is that all the different possible accounts, with variety of transfers, can be thought about in terms of information. Or in other words, information is what seems to be transferred here.

Information theory, as is used currently, is based on the work of Claude Shannon (1948), who described the channel through which the information is transmitted as consisting of a transmitter and a received, through which the signal is transmitted, and there is noise that interferes with transmission. Shannon's work allowed quantification of information and described the idea of information transmission, and information loss during the transmission. Essentially to how information is used here, Shannon's theory allows for measuring the information independently from its meaning.

Another essential development in the information theory was the algorithmic information theory (AIT) based on Kolmogorov's work (1963). AIT allowed formalizing the notions of complexity and compressibility, in a nutshell the more complex a thing is, the longer description is necessary and a complete description of a car will be longer than the description of a rock. However, it is possible to describe those things, like a car and a rock, in comparable ways, using a chain of yes and no questions (e.g. is it manmade? No, is it natural? Yes, etc.). Compressibility allows then for cutting as much of the information without at the end losing any of its meaning (e.g. what is the shortest possible description of a car?).

With its dynamic developments:

Information theory, including other formal measures of information, is now very important to many sciences, including at least physics, bioinformatics and cognitive science. What makes this possible is that information theory gives us a very general formal framework that can be used to represent and assess any kind of process. Anything can be described informationally, from a person to a supernova to a tsunami. The formal framework of information theory ensures that the description, in spite of its unprecedented generality, is not empty or vacuous. (Illari, Russo, 2014, p.137)

But can the information theory be applied to something else than science, or if it can be a theory in humanities with a strong basis in mathematics? My intuitions here follow those of Illari and Russo, that it can. What however needs to be emphasized is that it can only be a good account of mental causation, one that does not falls back onto the old issues of causal exclusion, is when it is connected with the complex cause solution I have proposed above.

Three major strands of information theories can be identified (Adriaans and van Benthem, 2008):

1. Knowledge, logic, what is conveyed in informative answers /the world of epistemic logic and linguistic semantics.
2. Probabilistic, information-theoretic, measured quantitatively/ Shannon information theory.
3. Algorithmic, code compression, measured quantitatively /Kolmogorov complexity, liked to the foundations of computation.

The first information type can be described as:

The typical logic-based setting lets an agent acquire new information about what the real world is like, through acts of observation, linguistic communication, or deduction. A simple example would be an agent asking a question, and learning what things are like from an answer. Thus, three features are crucial: agents which represent and use the information, dynamic events of information change, and 'aboutness': the information is always about some relevant described situation or world. Here, we measure quality of information qualitatively in terms of new things agent can truly say: a quantitative measure may be handy, but it is not required. Finally, the formal paradigm for the theory is mathematical or computational logic. (Adriaans and van Benthem, 2008, p.11)

The first type is then, more than quantitatively assessing the mount of physical information, focused on the meaning of the utterances, contrary to the theory of Shannon. Moreover, the first approach specifically involves mentioning of the agent in his interaction with the world. Additionally the statements are describable in terms of logic, which allows for a formalization of it.

The second type is described by Adriaans and van Benthem, as:

By contrast, the typical Shannon scenario is about a source emitting signals with certain frequencies, say a 'language' viewed as a global text produces, and the information which a receiver picks up from this measured in terms of expected reduction of uncertainty. […] No *specific agency seems involved here, but the scenario does analyze major features of communication which are absent on the logical approach, such as probability of signals […]Finally, mathematical paradigms for the theory are probability theory and physics.* (Adriaans and van Benthem, 2008, pp.11-12)

This type is how information usually seems to be discussed. It allows for quantitative transfer of the information, but omits any meaning, and an agent. However the two approaches combined, could have the potential to explain both, the physical transfer of information, and the impact of the 'abstract' side of information processing, namely the meaning of sentences.

The third type, linked to the Kolmogorov's complexity, is not of immediate importance here, as it covers, like the second type, the physical side of information processing.

Therefore the information transfer account we would be after, that might have the potential to explain the type of causation that takes place when intentional action is generated. The cause would consist of the mental linguistic semantics, the abstract information we extract from the world, and the physical part, transmitted through brain, possibly somehow similar to Shannon's.

First accounts of information causality.

The first philosopher who offered an account of information causality, still only as physical causality, is John Collier, and his most basic idea is that

[…] causation is the transfer of a particular token or quantity of information from one state of a system to another. Physical causation is a special case in which physical information instances are transferred from one state of physical system to another. (Collier, 1999, p.215).

His approach, closely related to Shannon's, treats information again in physical terms, and focuses on quantitative, and not qualitative analysis of information transfer.

Collier also thinks that his approach could account for the types of causal interactions that are the interest of cognitive scientists. This further supports my intuitions here. Collier describes the current views of cognition as;

The most dominant current view of cognition is the syntactic computational view, which bases cognitive processes on formal relations between thoughts. Whether or not the theory is true, it shows that psychologists are willing to take it for granted that form (viz., the syntax of representations) can be causal. Fodor (1968) argues that the physical embodiment of mental processes can vary widely, if the syntactic relations among ideas are functionally the same. To understand the embodiment of mind, if we accept that cognitive processes derive their formal relations from underlying dynamics, we need an account of the role of form and information and dynamics. (Collier, 1999, p.218)

Collier refers to multiple realization here, stating that the causes can be realized differently and the semantics, or the form, can be causal. This follows the standard approach, where the mental and the physical, and in this case the form and the syntax are seen disjunctively, as, somewhat, competing causes. Since Collier's approach builds on Salmon's original account, he describes the information flow in terms of identity of information (Illari and Russo, 2014, p.139), where:

P is a causal process in system S from time t_0 to t_1 is some particular part of information S involved in stages of P is identical in t_0 and t_1. (Collier, 1999, p.222)

In his later work Collier (2011) describes an information channel as a family of isomorphisms (also discussed by Dretske, 1999, and Floridi, 2010). Illari and Russo (2016) describe the covariance of isomorphisms as:

[…] two systems a and b are couples in such a way that a's being (or type, or in state) F is correlated to be being (of type, or in state) G, then such correlation carries for the observer of a the information that b is G. For example, the dishwasher's yellow light (a) flashing (F) is triggered by, and hence is informative about, the dishwasher (b) running out of salt (G) for an observed O, like Alice, informed about the correlation. (Illari and Russo, 2016, 260)

Importantly, the way isomorphisms are defined by Illari and Russo involves the agent, who is informed, by the already existing correlation between the flashing light of the dishwasher and low salt levels. As we have seen previously in Shannon's case, and presumably also in Collier's himself, no actual agent is taken into account, as information causation is only analyzed as physical causality. In both cases, Illari and Russo's and Collier, isomorphism entails the existence of two systems, each consisting of a set of objects, and each of the objects, a set of attributes. Isomorphism is when attributes of one objects tells us something about attributes of another objects. However, when we look at the way the agent is informed, it seems that the cause of his reaction, e.g. adding more salt to the dishwasher, consists of both: (1) the mental knowledge, or a belief that the flashing light signifies the low level of salt, and (2) the physical flashing light.

The most recent approach of Collier can be defined as:

P is a causal connection in a system from time t_0 to t_1 if and only if there is a channel between s_0 and s_1 from t_0 to t_1 that preserves some part of the information in the first state. (Collier, 2011, pp.10-11).

Therefore a causal connection is defined by the informational identity at different stages of the channel of information.

Illari and Russo (2016) argue that the informational account of causation can help us understand how science explains the causal structure of the world. They integrate this account with mechanistic approach to causality (which I explained in the chapter on causality). What they ultimately look for is an account that combines the, production, information causality account, with a difference- making approach, however it seems that their approach is mainly focused on explanation, rather than what causality actually is.

One of the reasons for advocating information transfer as the type of causality that could be applicable to the mind-brain debate broadly, and free will more specifically, is the very broad application of the notion of information, information processing and transfer in neuroscience. One position is especially of interest here, the argument by P.U. Tse, who links it with a solution to the mental causation problem, and through it free will.

Tse's criterial causation as information causation account.

Most popular approaches in neuroscience, that investigate, directly or indirectly, the problem of free will, focus on action initiation, and voluntary action, and use rather high level processing to try and determine if free will is possible. An entirely different angle of the problem is investigated by Peter Ulric Tse (2013) who, unlike Libet, treats the problem of free will not as a problem of the specific causes for motor action, but as an instance of the broader mind-body problem. Moreover, he thinks that the whole process, starting from the acquisition of a new information, can be seen as part of the decision making. The main focus of the approach thus, does not lie in the causes of specific motor actions, but rather, in the possibility of the mind's causal influence on the world. Additionally, also unlike Libet, Tse extends his understanding of voluntary action to include not only overt but also internal actions, thus requiring no muscular activation (e.g. intentional shifting your attention from one object to the other).

Tse starts with the idea that free will might not in fact be an event but rather a "durationally extended process" (Tse, 2013), constructed from various (I assume both conscious and unconscious in this case) sub-processes. The immediate questions here are: (1) where does this process start, and it seems it is in fact the acquisition of some information, but this can be very vague, (2) if the sub-processes are not conscious how can they be free? The latter is problematic, since, if we want to say that only the relevant parts need to be conscious, we would have to be able to give a criterion of relevance of those. However, the approach leaves space for the causality of intentions, but also runs a risk of stating that the whole brain is relevant for an intention and behaviour, which, while true, is also uninformative. The will is defined by Tse as:

[…] the agent's choice or decision to perform or not to perform an action, either later (via formation of distal intentions) or right now (via proximal intentions). Will is the capacity to so choose or decide to act. (Tse, 2013, pp. 20).

Unlike Libet's definition, where free action is taken to *(a) […] arises endogenously, not in direct response to an external stimulus or cue; (b) there are no externally imposed restrictions or compulsions that directly or immediately control the subject's initiation and performance of the act; and (c) most important, subjects feel introspectively that they are performing the act on their own initiative and that they are free to start or not to start the act as they wish.* (Libet, 1985, p.529-530)

Tse assumes that the possibility of free will requires an account of mental causation, but the latter could be supported on a very basic neuronal level, in the initiation of what will become the overt action. However, similarly to how it is done in metaphysical discussions on free will, Tse discussed the problem with the determinism/indeterminism debate. For Tse the mental would have a causal effect already in the initial stages of the process resulting in action, therefore the current experimental approaches do not look in the right place. Here the mental would be causally efficacious already at the neuronal level, and the possibility of this is explained by the core idea of his approach, criterial causation.

Tse argues that criterial causation not only explains how mental causation would be possible, but also solves one of what he considers the core problems, namely the impossibility of self-causation (the idea that our decisions are influenced by our motives, intuitively the core of the idea of mental causation which is needed if we consider ourselves as the choosing agents, who cause overt actions in the world). Contrary to Tse, I do not think that self-causation is actually the core of the problem, however, since it is intimately linked to the mental being causal (as 'us' that Tse thinks about is a mental event), therefore I will treat this as synonymous.

The criteria that form a basis for the account are a:

[…] set of conditions on input that can be met in multiple ways and to differing degrees. (Tse, 2013, p.22).

As such, the criteria considered here are similar to what Haggard considers alternative possibilities, where:

Movement neuroscience understands 'could have done otherwise' in terms of a process of action selection, that arbitrates between competing action alternatives, and selects the one that eventually occurs. […] if we have evidence from brain measurements that two different *action alternatives were developed in parallel, up to a late stage just before action execution, and if we can identify the particular brain process that led to one action alternative being selected and the other deselected, we may sat that the person, or animal, could have done otherwise.* (Haggard, 2011, p.9).

The main difference between what Haggard argues, and criterial causation, seems to be the initiation of such action selection processes, where for the former seems to assume such processes come into play in late stages of processing, while Tse seems to suggest that those selections happen already on a very basic level. On his account neurons assess incoming information based on the criteria and, if met, the neurons fire.

The neuronal model of mental criterial causation has three distinct stages:

(1) *new physical/informational criteria are set in a neuronal circuit on the basis of preceding physical/mental processing at time t1 , in part via a mechanism of rapid synaptic resetting that effectively changes the inputs to postsynaptic neuron*,

(2) *at time t2, inherently variable inputs arrive at the postsynaptic neuron*, and

(3) *at time t3 physical/informational criteria are met or not met, leading to postsynaptic neural firing or not*. (Tse, 2013, p.133).

The physical realization of informational criteria happens on the neuronal level, where the information from the input triggers the neuron to fire, if such criteria are met. Neurons transform, and communicate the information by changing individual action potential spikes into spikes that are being transmitted to other neurons, and place criteria for firing that can be met only by a specific subset of information, so only that subset can trigger the neuron to fire. Mental events can, on this account of causation, cause future mental and physical events by changing the criteria for firing used by neurons for the future inputs.

In simpler terms, neurons react on the basis of certain criteria, which can be described as conditions for firing, and only if those conditions are met the neuron will react and fire. Those conditions are met or not because of the information coded somehow in the input entering the brain. According to Tse, new information entering the brain contained in a new input can change the previous criteria according to which the neurons are firing. After the criteria are changed, the next time a new input will arrive, those new changed criteria will have to be met for the neuron to be activated. For simplicity, I will use the same example as the one discussed by Tse, namely the one referring to Orion. Before we have any knowledge about the pattern of stars in the constellation of Orion, what Orion is, or know how the constellation of Orion looks like, no neurons fire when the pattern is presented as an input. However, once the criteria (i.e. the input looking like the constellation of Orion) is set, the next time we will see that pattern, a specific neuron will fire. In the example of Orion, it is clear that Tse wants to say that both the physical realizer of the Orion, as well as the abstract meaning, stored somehow in the memory, trigger the physical reaction of the neuron. What he would probably not state, is that the two compete with one another, in the standard sense of mental and physical causal impacts.

The pattern of information is multiply realizable, meaning that the neuron will fire as a response to a certain pattern, in this case the pattern of Orion, regardless of it being presented in stars, grains of sand or fruits. The information is, in fact, the pattern, so the transfer of information avoids the issue of energy transfer. This could suggest that Tse's approach is not one of production, but difference-making, but here I want to argue that information account would be more appropriate. Mental events themselves are, for Tse, realized in physical events and they can be causes of both mental as well as physical events through initiating a change in the current state of the neuron, its criteria for firing. If we define a mental event through its informational content, the causality from one system (neuronal firing criteria pre introduction of the information), to the other (neuronal firing criteria post introduction of the information) is clearly an example of the informational causality, where the same information is preserved throughout the whole process.

Promising as it is, Tse's account has many problems. Firstly the notion of realization he operates with is unclear. As we have seen in the previous chapters, this is not an uncommon problem, however Tse treats the notion as if intuitively clear, when he argues:

[…] even though mental events are realized in physical events they (i.e., their physical realization) can cause subsequent physical and mental events by preparing new decoders, or changing the criteria for firing on already existing decoders. (Tse, 2013, p. 25).

It is unclear if Tse assumes the events are causal through their physical realizers or they are causal because of the mental side of the event. As we have seen already, the issue is essential, since, if the mental is only causal because of its physical realizer, it is epiphenomenal. The same happens if we think that all the causal powers at play come from the physical brain activation, then again the mental, while not reducible to the physical, is also epiphenomenal.

Secondly, Tse bases his solution of the mental causation problem, and through that the free will problem, on a specific understanding of the causal exclusion argument and causation itself. He assumes that the main issue for mental causation is the impossibility of self-causation, which I argue is not the case. Moreover, he links the problem of mental causation with the problem of compatibility of free will with the deterministic nature of the universe, which I specifically avoid here.

If the main obstacle was in fact self-causation Tse's argument would have shown how to escape it only on a very basic level, however usually the problem of self-determination is connected to conscious decisions rather than biological unconscious processes. If on the other hand we assume, as I do, that the problem lies elsewhere, namely in the overdetermination principle being at odds with the causal closure principle (the notion that all physical events have sufficient physical causes), Tse's approach does not seem to solve the issue at hand. If we are not willing to accept dualism nor reduction, we need to accept some form of coexistence of mental and physical, and the most widely used is the relation of supervenience, which puts into question the causality of mental qua mental and not qua physical, as put forward by Jaegwon Kim (e.g. 1998, 2005).

Since the mental events are always also physical (as they have the physical basis), in order to show how mental causation would work, we would have to show that what is the sole cause of another event is the mental part and not its physical basis. It has been discussed over and over here, that this is very challenging, if we try to sustain that the physical also enters the causal relation here, and given the causal closure of the physical, which states that all physical events have sufficient physical causes, and the lack of overdetermination (except from the rare cases of genuine overdetermination), the mental is effectively excluded from the causal chain. This might change if we treat them both as one cause. Tse in his argument, however, since the explanations how mental and physical are related within his account, and why the input would be causal through its mental properties rather than the physical, are missing, does not solve the puzzle of causal closure. It might be the case that Tse's approach is closer to a reduction in which mental events are seen as identical to physical events, rather than to any form or nonreductive physicalism.

Also problematic is Tse's usage of terms like, already mentioned, deciding or/and choosing, which puts into question the role played by neurons in his scenario. The notions connected to decision making processes, are usually attributed to people, or agents, rather than circuits in the brain (discussed in Bishop, 2014). We tend to assume that an agent makes a decision, and the decision itself is a more complex computation, not completely reducible to the lower level, neuronal computation. On the other hand, Tse seems to attribute those, usually high-level computations, to neurons rather than individuals, which could suggest an underlying reduction of agents to brain circuits, which in turn makes the possibility of mental causation even more questionable. We don't need to fall into dualism in order to defend the efficacy of mental events (again we can agree on a form of nonreductive physicalism), but reducing those deeply human abilities to neuronal assessment of the incoming information seems to leave very little room for what we consider necessary for free will. It is unclear how neurons would even be able to assess the incoming information and choose or decide beyond a simple mechanistic, biological reaction, which is not what we have in mind neither for free will nor for mental causation. From the description of how the criterial causation is supposed to work, it seems that neurons only possess the ability to check if the input information fulfils the conditions that would make the neuron fire or not, and not much beyond that. This again is not what we expect from a free action, and we tend to think that what is needed is more of a high level property of the system, not possible on this very basic level. It is very hard to see how neurons would "decide" and decode meaning of abstract notions. Additionally, it leaves the processes underlying free will to be largely performed unconsciously, which stands in the opposition to an argument that is almost universally accepted - that free will requires consciousness and our free decisions or actions are done consciously. If all mental causation would be limited to the, unconscious, resetting of the neuronal criteria, it would put into question what kind of control can we have on the process and how does it differ from pure randomness. So although mental causation may work on that very low computational level, it also has to be causal somewhere later in the decision making processes, directly influencing our intentions or decision.

Another problematic angle of Tse's argument lies in the notion of information, which, beyond its connection to the notion of input pattern, is never precisely defined. Moreover, it is likely that, if information should be causal, and play a role for free will, the influence the information has should be, at least the first time around, when the criteria are changed, somehow conscious (as we have seen on Illari and Russo's example of isomorphism, the agent recognizes the information that the blinking light in the dishwasher is transmitting). The automatic changing of the neurons' firing criteria as the basis for free action seems insufficient and resembles more a computer program executing a task, rather than a basis for a free action. It is also unclear what is the role played by neurons in the information assessment, and through the usage of notions like deciding, the complete mappability of agenthood onto the neuronal activity is suggested. This puts the concept of the agent as well as the importance of higher level computation into question. If everything about the human action would then be just a flow of energy between neurons, the opposite of what Tse would like to argue for (further discussed in Bishop, 2014) would potentially be proven - the mental is reducible to the physical and even if it isn't, it's epiphenomenal (plays no causal role in the cause-effect chain). Tse either implicitly allows reduction, when he states that everything is just particles interacting with other particles (Tse, 2013), or he just leaves the mental as being epiphenomenal.

There are however great advantages to Tse's account, and they come from philosophical reflections on the nature of free will. First of all, Tse recognizes the extended in time nature of our decision making processes resulting in that result in an overt action. He tries to build an account in which free will is given on such a basic level as to possibly account for all the necessary steps in making a choice. However, this should be connected with an account of how mental causation could additionally be used on higher levels of computation, where it could be connected with consciousness, and not only on the neuronal one, where it could be connected with consciousness. This, potentially, might entail adding either different levels of computation or accepting the notion of emergence. Then mental causation will potentially happen in two different points of the process leading to action, both fulfilling a very different role.

The other essential notion is the one of information. I have already mentioned some of the problems with how the notion is used in Tse's account, yet the advantages of both information as well as the idea of patterns should not be overlooked. It has been proven by neuroscientists that the human brain remains plastic throughout most of our lives (neuroplasticity), which undermined the previous knowledge that it develops rapidly during childhood and then remains static. Changes in the structure of the brain can be caused by physical events but potentially also by all interactions and cultural patterns with which we come into contact. All of them possibly have impact on the physical structure of our brain and through that on the decisions we are making, or at least those are the intuitions we have when we think about what directs our choices. Tse, through the notions of information and pattern, offers a possibility to explain how exactly that happens. The solution is not without its problems as we have seen before, however it might be a first step in a new research direction.

Conclusions.

In this, concluding chapter of my dissertation, my aim was twofold. Firstly, after establishing that causally efficacious mental is necessary for the possibility of free will, and that causal exclusion poses a serious threat to free action, I have analyzed different ways that have been used to avoid or solve the causal exclusion problem. I have focused on two manners: changing the way we define causality (from production, which was the approach Kim used), and changing what we consider to be the cause, through rejecting one of the premises of the argument (the no-overdetermination principle). I have argued that all these attempts fail to address the problem. I have then argued that, for the purposes of the free will debate, it might be necessary to both change the notion of a cause and how we think about causality (to accommodate the changed notion of the cause). I have hypothesized that the mental and the physical cannot be taken to be two separate causes, as they are not only not independent, but it seems that, based on the discussions in neuroscience, and folk intuitions, none of them seems truly sufficient in producing the intentional action. This does not mean that the physical causal closure would be broken, the physical cause is sufficient in producing the physical action, but to produce intentional action, both the physical and the mental are necessary, much like in INUS conditions described by Mackie. Therefore, in the specific case of voluntary, or intentional, action both the physical and the mental are necessary, and moreover it seems to make more sense to talk about one cause with both physical and mental properties. The approach where we take them to be two separate, competing causes, seems to be the source of the failure of all the approaches discussed. Moreover the physical and the mental are defined in different ways: one ontologically, the other functionally, which makes me question if it is not counterproductive to consider them as competing and not complementary.

I argue that, in order to explain the dynamics between the new approach to causes, a new approach to causality is also needed, and I hypothesize that such work could be done by information account of causality, which has the potential to incorporate the idea of one cause with two different properties (as in information the physical realization, and the meaning of information). Since the approach is new, and based on few, underdeveloped positions, here I only attempt to lay the foundations of an approach, which has to be further developed, as future work. I have described the few existing positions, starting from the basic approach of Shannon, through Collier's first attempt to build an informational causal approach, Illari and Russo's suggestions that it could be incorporated into a mechanistic understanding of causality, to Tse's criterial causation, which I consider to be an example of information causation in neuroscience.

Since all these regard the physical transfer of information, and is focused on physical causality, further work requires developing an account on how the informational signal is related to the meaning of the information. According to Dretske:

The information embodied in a signal (linguistic or otherwise) is only incidentally related to the meaning (if any) of that signal. Typically, of course, we communicate with one another, exchange information, by exploiting the conventional meaning of signs. We convey information by using signs that have a meaning corresponding to the information we wish to convey. But this practice should not lead one to confuse the meaning of a symbol with information, or amount of information, carried by the symbol. (Dretske, 1999, p.44).

Therefore the task is daunting, as not only we have to explain how the meaning works inside the information, as its parts, but additionally we should avoid mistaking symbols with either one of them. Especially since the meaning of a symbol could in fact be different than the information itself. Therefore, a clear distinction should be drawn between the symbolic representations of meaning, and the notion of information. Also the signals, which are usually used in the descriptions of transfer of information, carry information, but might have meaning. Again, according to Dretske:

What information a signal carries is what it is capable of 'telling' us, telling us truly, about another state of affairs. Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it. If everything I say to you is false, then I have given you no information. (Dretske, 1999, p.44)

This again reiterates Dretske's view that information has to be true, while meaning does not, which again could influence future theories of information causation.

The main issue here seems to lie in the difference between the study of conditions of communication and particular messages, which are the vehicles of information (and meaning), and the informational content of the message seems to rely on what we can learn from it (Dretske, 1999), which again brings back the notion of an agent. However, as we have seen in the previous chapters, the notion of the agent itself is problematic for the free will debate (the problem of a disappearing agent). Moreover, it seems that said agent already has to be familiar with the possibilities, when it comes to what information can be transmitted. If we take Illari and Russo's example of a dishwasher, it seems that the information the salts are finished, requires previous knowledge of what that blinking light can mean. This could be connected to learning changing criteria on the neurons firing in response to the blinking light (here I am only speculating).

Therefore further work is needed to develop such account, but I am convinced that such position can be developed, and address the issues

The overall aim and summary.

The discussions on the notion of free will currently seem to have ended up in a deadlock. The discussions are mainly focused on the problem of determinism and its threat to the possible existence of free will. However, without a conclusive answer regarding the nature of the universe, it will be difficult to move from this puzzle.

Here I have argued that, while determinism is a threat, it should be decoupled from the notions of predictability and causality, as both of them come with their own set of problems. I have chosen causality here, and I have argued that, in fact, one of the conditions of necessity for the possibility of free will is having the right type of causes. Those causes, in all discussions, are talked about as mental, which agrees with our own experiences of agency, and the sense of it. When framed like this, the main threat to free will has nothing to do with the deterministic nature of the universe but rather with the causally inefficacious mental. If our mental events do not cause anything we do not have free will. However, if they do, free will is also not guaranteed, as this problem is challenging and multilayered.

I have shown how the standard solutions to the causal exclusion, the main threat to mental causation, fail to account for the causally efficacious mental, and especially they fail to account for the type of causes we need for free will. Therefore, in the way philosophers usually do, I have decided to rethink the core concepts, and I reached a conclusion that most of the problems here, stem from our notion of a cause. And, that besides my own intuitions, there are other good reasons to believe that it might be wrong, or not reflecting our intuitions and the ways in which we use it. I have proposed a new approach to the notion of the cause, where there is only one, instead of two competing causes. Said one cause has mental and physical properties, that together cause a behavior, but are not taken to compete, and their causal impact on the action is not distinguishable. They necessitate one another, and, while there are still many questions about their relationship, and how the mental is in fact realized by the brain, they would not be related to the question of causality anymore.

I have also proposed a very brief sketch of a new approach to causality, that could have a potential to accommodate such complex causes. The causality as information transfer account has been gaining popularity in the literature on causality, but for now, it is usually discussed in connection to physical causation solely. This concluding chapter could serve as the initial fundament to build a full position, in line with the way neuroscientists talk about the processes going on in the brain.

List of Tables :

List of Figures :

References

Adriaans, P. van Benthem, J. (eds.)(2008). *Philosophy of Information.* Elsevier.

Aguilar, J.H., Bukareff, A.A. (eds.) *Causing Human Actions: New Perspectives on the Causal Theory of Action.* MIT Press

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology,* 32, 665-683.

Alexander, P., Schlegel, A., Sinnott-Armstrong, W., Roskies, A. Wheatley, T., Tse, P.U. (2016). Readiness potentials driven by non-motoric processes. *Conscious Cognn.* 39, 38-47.

Alvarez, M., Human, J. (1998). Agents and their Actions. *Philosophy,* 73(2), 219-245.

Alvarez, M. (2013). Explaining Actions and Explaining Bodily Movements. In: D'Oro, G., Sandis, C. (eds.) *Reasons and Causes: Causalism and Anti-causalism in the Philosophy of Action.* Palgrave Macmillan, 141-160.

Andersen, Holly (2006) Two Causal Mistakes in Wegner's Illusion of Conscious Will; available at: http://philsci-archive.pitt.edu/3008/1/Andersen_Two_Causal_Mistakes.pdf

Andersen, R.A., Buneo, C.A. (2002). Intentional maps in posterior parietal cortex. *Annu. Rev. Neurosci*, 25, 189-220.

Anderson, J. (1938). The Problem of Causality. *Australasian Journal of Psychology and Philosophy*, 16 (reprinted in Studies in Empirical Philosophy, Sydney, 1962)

Andersen, R., Buneo, C.(2002). Intentional Maps in Posterior Parietal Cortex. *Annual Review of Neuroscience,* 25, 189-220

Anscombe, G.E.M. (1957). *Intention*. Basil Blackwell. [reprinted 2000 Harvard University Press]

Anscombe, G.E.M. (1971). *Causality and Determination: An Inaugural Lecture*. Cambridge University Press.

Árnadóttir, S.T., Crane, T. (2013). There is no Exclusion Problem. In: Gibb S.C., Lowe E.J., Ingthorsson R.D. (eds.). *Mental Causation and Ontology*, Oxford University Press, 248-266.

Bandura, A. (1982). Self-efficacy in human agency. *American Psychologist,* 37, 122-147.

Banks, W.P., Isham, E.A. (2009). *Psychological Science*, 20(1), 17-21.

Baumeister, R.F., Masicampo, E.J., DeWall, C.N. (2009). Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness. *Personality and Social Psychology Bulletin,* 35 (2), 260-268.

Baumgartner, M. (2009). Interventionist Causal Exclusion and Non-reductive Physicalism. *International Studies in the Philosophy of Science* 23, 161-178.

Baumgartner, M. (2012), The Logical Form of Interventionism. *Philosophia* 40, 751-761.

Baumgartner, M. (2013). Rendering Interventionism and Non-reductive Physicalism Compatible. *Dialectical* 67, 1-27

Baumgartner, M., Gebharter, A. (2016), Constitutive Relevance, Mutual Manipulability, and Fat-Handedness*, The British Journal for the Philosophy of* Science 67, 731-756.

Baumgartner, M. (2017), The Inherent Empirical Underdetermination of Mental Causation*, The Australasian Journal of Philosophy,* doi: 10.1080/00048402.2017.1328451

Bayne, T.(2011). Libet and the Case for Free Will Scepticism. In *Free Will and Modern Science* (Swinburne, R. ed.) British Academy.

Baysan, U. (2015). Realization relations in metaphysics. *Minds and Machines.* Available at:

eprints.gla.ac.uk/105815/1/105815.pdf

Bechara, A., Damasio, H., Lee, G.P. (1999). Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision- Making. *The Journal Of Neuroscience*, 19 (13) 5473-5481.

Bechtel, W., Richardson, R. C. (1992). Emergent phenomena and complex systems. In Beckermann, A. Flohr, H. , Kim, J. (Eds.) *Emergence or Reduction? Essays on the prospects of nonreductive physicalism*, Berlin: Walter de Gruyter Ver lag, 257-288.

Beebee, H. (2013). *Free will: An Introduction.* Palgrave Macmillan.

Bennett, J. (1988). *Events and Their Names.* Hackett.

Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Nous*, 37, 471–497.

Bernstein, S. (2016). Overdetermination Undetermined. *Erkenntnis*, 81 (1), pp. 17-40.

Bishop, R.C. (2014). Review of the Neural Basis of Free Will, available at: http://ndpr.nd.edu/news/neural-basis-of-free-will-criterial-causation/

Blakemore, S-J., Sirigu, A. (2003). Action prediction in the cerebellum and in the parietal lobe. *Experimental Brain Research,* 153 (2), 239-245.

Block, N. (2003). Do Causal Powers Drain Away? *Philosophy and Phenomenological Research*, 67 (1), 133-150.

Blondeau, J., Ghins, M. (2013). Is there an intrinsic criterial for causal lawlike statement? *International Studies in the Philosophy of Science* 26 (4), 381-401.

Bode, S., Hanxi He, A., Soon, Ch.S., Trampel, R., Turner, R., Haynes, J.D. (2011). Tracking the Unconscious Generation of Free Decision Using Ultra-High Field fMRI. *Plos One.* Available at:

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021612

Brass, M., Haggard, P. (2007). To do or not to do: the neural signature of self-control. *J. Neurosci.,* 27 (34), 9141-5

Brass, M., Haggard, P. (2008). The what, when, whether model of intentional action. *Neuroscientist*, 14 (4), 319-325.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Campbell, J.K., O'Rouke, M., Shier, D. (2004). Freedom and Determinism: A Framework. In: Campbell, J.K., O'Rouke, M., Shier, D. (eds.) *Freedom and Determinism*. A Bradford Book. The MIT Press, 1-19.

Campaner, R., Galvanotti, M.C. (2007). Plurality in causality. In: Machamer, P., Wolters, G. (eds.) *Thinking about causes. From Greek Philosophy to modern physics*. University of Pittsburgh Press, 178-199

Carruthers, P.(2007). The Illusion of Conscious Will. *Synthese* 159, 197-213

Churchland, Patricia (1986). *Neurophilosophy.* MIT Press.

Crane, T. (1995). The mental causation debate. *Proceedings of the Aristotelian Society* 69, 211-236.

Craver, C.F., Darden, L. (2001). Discovering Mechanisms in Neurobiology: The Case of Spatial Memory in: Machamer, P.K., Grush, R., McLaughlin, P. (eds.), *Theory and Method in Neuroscience*, Pittsburgh, PA: University of Pittsburgh Press. 112-137.

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press: Oxford.

Collier, J. (1999). Causation is the Transfer of Information. In: Sankey, H. (eds.) *Causation, Natural Laws and Explanation.* Dodrecht: Kluwer.

Collier, J. (2011). Information, causation and computation. In: Crnkovic G. D., Burgin, M. (eds.) *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*. World Scientific.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy,* 60(23), 685–700; reprinted in his 1980, *Essays on Actions and Events,* Oxford: Clarendon Press, 3–21

Davidson, D. (1980). Mental Events. In *Essays on Action and Events.* Oxford, Clarendon Press, 207-225

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking Adult.

Deecke, L. (2014) Experiments Into Readiness for Action: 50th Anniversary of the Bereitschaftspotential. *World Neurology. The official newsletter of the World Federation of Neurology,* 29 (3), 1, 6-9.

Dennett, D. (1984). *Elbow Room.* Cambridge, MA: MIT Press.

Dennett, D. (2003). *Freedom Evolves.* Viking Press

Dennett, D. (2013). Please Don't Feed the Bugbears. In Russel, P., Deery, O. (eds.) *The Philosophy of Free Will. Essential Readings from the Contemporary Debates.* Oxford University Press, pp.43-59.

Dennett, D. (2014). Seduced by Tradition. Comments on Gazzaniga. In Sinnott-Armstrong, W. (eds.) *Moral Psychology. Volume 4: Free Will and Moral Responsibility*. A Bradford Book, The MIT Press, pp. 75-80.

Descartes (1641). *Meditation 6*, available here:

https://en.wikisource.org/wiki/Meditations_on_First_Philosophy/Meditation_VI

Desmurget, M., Grafton, S. (2000). Forward modelling allows feedback control for fast reaching movements. *Trends Cogn. Sci.*, 4, 423-431.

Desmurget, M., Reilly, K.T., Richart, N., Szathmari, A., Mottolese, C., Sirigu, A. (2009a). Movement initiation after parietal cortex stimulation in humans. *Science*, 324 (5928), 811-813.

Desmurget, M., Sirigu, A. (2009b). A parietal-premotor network for movement intention and motor awareness. *Trends Cogn. Sci.* 13 (10), 411-419.

Dijkstra, N., de Bruin, L. (2016). Cognitive Nuroscienceand Causal Inference: Implications for Psychiatry. *Front Psychiatry*. https://doi.org/10.3389/fpsyt.2016.00129

D'Oro, G., Sandis, C. (eds.) *Reasons and Causes: Causalism and Anti-causalism in the Philosophy of Action*. Palgrave Macmillan, 141-160.

Dowe, P. (1992). An empiricist defense of the causal account of explanation. *International Studies in the Philosophy of Science,* 6 (2), 123-128.

Dretske, F. (1977). Refering to events. *Midwest Studies in Philosophy,* 2 (1), 90-99.

Dretske, F. (1999). *Knowledge and the Flow of Information*. Cambridge University Press.

Edelman, G. (1989). *The Remembered Present.* New York: Basic Books.

Frankfurt, H.G. (2013). Alternate Possibilities and Moral Responsibility. In Russel, P., Deery, O. (eds.) *The Philosophy of Free Will. Essential Readings from the Contemporary Debates.* Oxford University Press.

Feigl, H. (1958) The "Mental" and the "Physical". In Feigl, H., Scriven, M., Maxwell, G. (eds.), *Concepts, Theories and the Mind-Body Problem* (*Minnesota Studies in the Philosophy of Science, Volume 2).* University of Minnesota Press.

Floridi, L. (2010). *Information. A Very Short Introduction.* Oxford University Press.

Fried, I., Mukamel, R., Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*. 69 (3), 548-562

Fodor, J.A. (1968). *Psychological Explanation*. Random House.

Fodor, J.A. (1974) Special Sciences: Or the Disunity of Science as a Working Hypothesis. *Synthese*, 28: 97–115

Garrett, D. (2009). Hume. In: Beebee, H., Hitchcock, Ch., Menzies, P. (eds.) *The Oxford Handbook of Causation*. Oxford University Press.

Gasking, D. (1955). Causation and Recipes. *Mind*, New Series, Volume 64 (256), 479-487.

Gazzaniga, M.S. (2014). Mental Life and Respoonsibility in Real Time with a Determined Brain. In Sinnott-Armstrong, W. (eds.) *Moral Psychology. Volume 4: Free Will and Moral Responsibility*. A Bradford Book, The MIT Press, 59-74

Ginet, C. (1966). Might We Have No Choice? In Keith Lehrer (ed.), *Freedom and Determinisim*. Random House.

Habermas, J. (2008). *Between Naturalism and Religion*. Cambridge: Polity.

Haggard, P., Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Exp. Brain Res.,* 126, 128-133.

Haggard, P., Libet, B. (2001). Conscious intention and brain activity. *Journal of Consciousness Studies*, 8(11), 47-63.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, 9, 934-946.

Haggard, P. (2011.) Does brain science change our view of free will? In Swinbourne, R (eds.) *Free Will and Modern Science*. British Academy.

Haggard, P.(2014). Free Will. In Mele, A. (eds.) *Sourrounding Free Will: Philosophy, Psychology, Neuroscience.* Oxford Scholarship Online.

DOI: 10.1093/acprof:oso/9780199333950.001.0001

Harbecke, J. (2008). *Mental Causation: Investigating the Mind's Powers in a Natural World.* Walter de Gruyter.

Harmon-Jones, E., Mills, J. (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology.* American Psychological Association.

Hart, H.L.A., Honore, A.M. (1959). *Causation in the Law.* Oxford University Press.

Heil, J. (2011). Powers and the realization relation. *The Monist*, 94(1), 34 – 53.

Horgan, T. (1997). Brute Supervenience, Deep Ignorance, and the Problem of the Many, *Philosophical Issues* 8, 229-36

Hume, D. (1739). *A Treatise of Human Nature http://www.davidhume.org/texts/thn.html*

Humphreys, P. (2008). Synchronic and diachronic emergence. *Minds & Machines*, 18, 431–442.

Illari, P., Russo, F. (2014). *Causality. Philosophical Theory Meets Scientific Practice.* Oxford University Press.

Illari, P. Russo, F. (2016) in Floridi, L. (eds.) *Routledge Handbook of the Philosophy of Information*. Routledge.

Kallestrup, J. (2006). The causal exclusion argument. *Philosophical Studies*, 131, 459–485.

Kane, R. (1989). Two Kinds of Incompatibilism. *Philosophy and Phenomenological Research,* 50, 219-254.

Kane, R. (1996). *The Significance of Free Will.* Oxford University Press.

Kane, R. (2002.) Introduction. In Kane, R. (Eds.) *The Oxford Handbook of Free Will.* Oxford University Press., pp.3-41.

Kane, R. (2011). Introduction. In Kane, R. (Eds.) *The Oxford Handbook of Free Will.* Oxford University Press, pp. 3-35.

Kim, J. (1984). Epiphenomenal and Supervenient Causation. *Midwest Studies in Philosophy*, 9: 257–70.

Kim, J (1987). 'Strong' and 'global' supervenience revisited. *Philosophy and Phenomenological Research* 48, 315-326.

Kim, J. (1988). Supervenience for multiple domains. *Philosophical Topics* 16 (1), 129-150.

Kim, J. (1985). Supervenience, determination and reduction. *Journal of Philosophy*, 82 (11), 616-618.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays.* Cambridge Studies in Philosophy. Cambridge University Press.

Kim, J. (1998). *Mind in a Physical World.* MIT Press.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kim, J. (2009). Mental Causation. In: McLaughlin, B., Beckermann, A., Walter, S. (eds.) *The Oxford Handbook of Philosophy of Mind.* Oxford University Press.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63, 190-193

Kolmogorov, A. (1963). On Tables of Random Numbers. Reprinted in: (1998) *Theoretical Computer Science.* 207 (2): 387–395.

Kornhuber, H. H., Deecke, L. (1965) Hirnpotentialanderungen bei Willkiirbewegungen und passivenBewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pfliigers Archivfur Gesamte Physiologie*, 284, 1 -17. (in English: Changes in the brain potential in voluntary movements and passive movements in man: readiness potential and reafferent potential)

Lafargue, G., Duffau, H. (2008). Awareness of intending to act following parietal cortex resection. *Neuropsychologia,* 46, 2662-2667.

Lau, H.C., Gleason, C.A., Wright, E.W., Pearl, D.K. (2007). Manipulating the experienced onset of intention after action execution. *J. Cognitive Neurosci.* ,19, 81-90.

Laplace, P.S. (1902). *A Philosophical Essay on Probabilities.* J. Willey; Chapman and Hill.

Leotti, L., Iyengar, S.S., Ochsner, K.N. (2010). Born to choose: The Origins and Value of the Need For Control. *Trends Cogn. Sci.,* 14 (10), 457-463.

Lepore, E., McLaughlin, B.P. (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davison.* Blackwell.

Lewis, D. (1973a). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1973b). Causation. *Journal of Philosophy.* 70, pp. 556–67.

Lewis, D. (1981). Are we free to break the laws?*Theoria*. 47(3), 113-121

Lewis, D. (1986). *The Plurality of Worlds.* Oxford Blackwell.

Libet, B., Wright, E.W., Gleason, C.A. (1982). Readiness-potentials preceding unrestricted 'spontaneous' *vs* pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology,* 54, 322-335.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences,* 8, 529-566.

Libet, B. (1999). Do we have Free Will? *Journal of Conscious Studies.* 6 (8-9), 47-57.

Libet, B. (2004). *Mind Time*. Harvard University Press.

List, Ch., Menzies, P. (2017). My brain made me do it: The exclusion argument against free will and what's wrong with it. In: Beebee, H., Hitchcock, C., Price, H. (eds.) *Making a Difference*. Oxford University Press.

Lowe, E. (1993). The causal autonomy of the mental. *Mind,* 102, 629–644.

Mackie, J.L. (1965). Causes and Conditions. *American Philosophical Quarterly,* 2 (4), 245-264.

Mackie, J.L. (1974, reprinted 1980). *The Cement of the Universe. A Study of Causation.* Oxford University Press

Machamer, P.K., Darden, L., Craver, C.F. (2000). Thinking about Mechanisms. *Philosophy of Science,* 67:1-25.

Marcel, A.(2003). The Sense of Agency: Awareness and Ownership of Action. In Roessler, J., Eilan, N. (eds.) *Agency and Self-Awareness.* Oxford Clarendon Press.

McLaughlin, B.P. (1995). Varieties of supervenience. In: Savellos, E., Yalchin, O. (eds.) *Supervenience.* Cambridge University Press, 16--59

Menzies, P. (1989). Probabilistic causation and causal processes: A critique of Lewis. *Philospphy of Science,* 56 (4), 642-663.

Menzies, P., Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science,* 44 (2), pp.187-203.

Menzies, P. (2003). The causal efficacy of mental states. In: Walter, S., Heckmann, H.D. (eds.) *Physicalism and mental causation. The metaphysics of mind and action*. Imprint Academic. Available at:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.3730&rep=rep1&type=pdf

Mele, A. (1992). *Springs of Action.* Oxford University Press.

Mele, A. (2003). *Motivation and Agency.* Oxford University Press.

Mele, A. (2006). *Free Will and Luck.* Oxford University Press.

Mele, A. (2007). Persisting Intentions. *Nous*, 41 (4), 735-757.

Mele, A. (2008). *Free Will and Luck.* Oxford University Press.

Mele, A.(2009) *Effective Intentions. The Power of Conscious Will.* Oxford University Press.

Mele, A. (2012). Another Scientific Threat to Free Will? *The Monist*, 95 (3), 422–440.

Mele, A. (2013). Actions, Explanations and Causes. In D'Oro, G., Sandis, C. (eds.) *Reasons and Causes: Causalism and Anti-Causalism in the Philosophy of Action*. Palgrave Macmillan, 160–174

Mele, A. (2014a). *Free. Why Science Hasn't Disprove Free Will*. Oxford University Press.

Mele, A. (2014b). Free Will and Substance Dualism: The Real Scientific Threat to Free Will? In: Sinnott-Armstrong, W. (eds.) *Moral Psychology. Volume 4: Free Will and Moral Responsibility.* A Bradford Book. The MIT Press. 195-207

Mill, J.S. *A System of Logic, Ratiocinative And Inductive.* Available at: https://www.gutenberg.org/files/26495/26495-pdf.pdf

Miller, J., Shepherdson, P., Trevena, J. (2011). Effects of Clock Monitoring on Electroencephalographic Activity: Is Unconscious Movement Initiation an Artifact of the Clock? *Psychological Science,* 22(1), 103-109

Miller, J., Schwarz W. (2014). Brain signals do not demonstrate unconscious decision making: An interpretation based on graded conscious awareness. *Consciousness and Cognition* 24, 12-21

Müller, V. A., Brass, M., Waszak, F., Prinz, W. (2007). The role of the preSMA and the rostral cingulate zone in internally selected actions. *NeuroImage*, 37, 1354–1361

Moore, M. (2010) Renewed Questions about the Causal Theory of Action. In: Aguilar, J.H., Bukareff, A.A. (eds.) *Causing Human Actions: New Perspectives on the Causal Theory of Action.* MIT Press, 27-44.

Moore, D. (2012). Causal Exclusion and Dependent Overdetermination. *Erkenntnis,* 76 (3)

319-335.

Moore, D. (2017). Mental causation, compatibilism and counterfactuals. *Canadian Journal of Philosophy*, 47 (1), pp. 20-42.

Moors, A. (2017). The integrated theory of emotional behavior follows a radically goal-directed approach. *Psychological Inquiry*, 28, 68-75.

Nagel, E. (1949) The Meaning of Reduction in the Natural Sciences. In Stouffer R.C. (eds.), *Science and Civilization*, University of Wisconsin Press, 99–135.

Nagel, E. (1961) *The Structure of Science. Problems in the Logic of Explanation*. Harcourt, Brace & World, Inc.

Nagel, E. (1970) Issues in the Logic of Reductive Explanations. In: Kiefer, H.E., Munitz K.M. (eds.), *Mind, Science, and History*, SUNY Press, 117–137.

Nahmias, E., Morris, S., Nadelhoffer, T., Turner, J. (2005). Surveung Freedom: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology* 18 (5), 561-584.

Nahmias, E. (2006). Folk Fears about Freedom and Resposibility: Determnism vs, Reduction. *Journal of Cognition and Culture* 6(1), 215-237.

Nahmias E., Coates J.D., Kvaran T. (2007). Free will, Moral responsibility and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy,* 31 (1), 214-242.

Nahmias, E., Shepard, J., Reuter, S. (2014a). It's OK if 'My Brain Made Me Do It': People's Intuitions about Free Will and Neuroscientific Prediction. *Cognition* 133(2), 502-516.

Nahmias, E. (2014b). Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences. In: Sinnott-Armstrong, W. (eds.), *Moral Psychology, vol. 4: Freedom and Responsibility*. MIT Press, 1-25.

Nahmias, E. (forthcoming). Your Brain as the Source of Free Will Worth Wanting: Understanding Free Will in the Age of Neuroscience. In: Caruso, G., Flanagan, O. (eds.), *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, Oxford University Press.

Nahmias, E., Deery, O. (forthcoming). Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood. *Philosophical Studies.*

Nanay, B. (2016). The role of imagination in decision-making. *Mind & Language* 31, 126-142.

Nichols S. (2006). Folk Intuitions on Free Will. *Journal of Cognition and Culture*, 6.1-2, 57- 86.

Nichols, S., Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science oof Folk Intuitions. *Nous*, 41 (4), 663-685.

O'Connor, T. (1995). Agent Causation. In: O' Connor, T. (Eds.) *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, New York: Oxford University Press, pp.173–200.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.

O'Connor, T. (2009). Agent-Causal Power. In: Handfield, T. (eds.), *Dispositions and Causes.* Clarendon Press, 189–214

O'Connor, T. (2011). Agent-causal theories of freedom. In Kane, R. (eds.) *The Oxford Handbook of Free Will.* Oxford University Press. 309-329.

Oshana, M. (1997). Ascriptions of responsibility. *American Philosophical Quarterly,* 34, 71-83.

Pacherie, E. (2008). The phenomenology of action: a conceptual framework. *Cognition*, 107 (1), 179-217.

Pacherie, E., Haggard, P. (2010). What are intentions? In: Nadel, L., Sinnott-Armstrong, W. (eds.) *Conscious Will and Responsibility. A tribute to Benjamin Libet*. Oxford University Press, 70-84

Pacherie, E. (2015). Conscious Intentions - The Social Creation Myth. In Metzinger, T., Windt, J.M. (Eds). *Open MIND*: 29(T). Frankfurt am Main: MIND Group, available at https://sites.google.com/site/pacherie/publications/topics

Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents.* Cambridge University Press, 3–36.

Paul, L.A. (2004). Aspect Causation. In Collins, J., Hall, N., Paul, L.A. (eds.) *Causation and Counterfactuals.* A Bradford Book. MIT Press.

Paulhus, D.L., Margesson, A. (1994). Free Will and Determinism (FAD) scale. Unpublished questionnaire. University of British Columbia, Vancuver, Canada (after Vohs and Schooler, 2008).

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press.

Pereboom, D. (1995). Determinism al dente. *Nous*, 29 (1), 21-45.

Pereboom, D. (2001). *Living without free will.* Cambridge University Press.

Pereboom, D. (2002). Robust nonreductive materialism. *Journal of Philosophy*, 99, 499–531.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life.* Oxford University Press.

Pockett, S., Purdy, S. (2010). Are voluntary movements initiated preconsciously? The relationships between readiness potentials, urges and decisions. In: Sinnott- Armstrong, W., Nadel, L. *Conscious Will and Responsibility: A Tribute to Benjamin Libet.* Oxford University Press.

Polger, T. (2004). *Natural Minds*. MIT Press.

Putnam, H. (1967). The Nature of Mental States. In: Putnam, H. (1975) *Mind, Language, and Reality*. Cambridge University Press. 362-385

Raatikainen, P. (2010) Causation, exclusion and the special sciences. *Erkenntnis* 73 (3), 349-363.

Raatikainen, P. (2013). Can the mental be Causally Efficacious? In: Talmont-Kaminski, K., Milkowski, M. *Regarding the Mind Naturally: Naturalist Approaches to the Sciences of the Mental.* Cambridge Scholars Press.

Rakos, R., Laurene, K.R., Skala, S., Slane, S. (2008). Belief In Free Will: Measurement And Conceptualization Innovations. *Behavior and Social Issues.* 17, 20-39.

Rigoni, D., Brass, M., Satori, G. (2010). Post-action determinants of the reported time of conscious intention. *Frontiers Hum. Neuoscience.* https://doi.org/10.3389/fnhum.2010.00038

Rigoni, D., Kuhn, S., Satori, G., Brass, M.(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), 613-618.

Roskies, A. (2006). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences*, 10: 419-423.

Roskies, A.(2011). Why Libet's studies don't pose a threat to free will. In Sinnott-Armstrong, W., Nadel, L. (eds.) *Conscious Will and Responsibility.* New York: Oxford University Press, 11-22.

Roskies, A.(2014). Can Neuroscience Resolve Issues about Free Will? In Sinnott-Armstrong, W. (eds.) *Moral Psychology. Volume 4: Free Will and Moral Responsibility*. A Bradford Book, The MIT Press, 103-126

Russel, B. (1948). *Human Knowledge: Its Scope and Limits*. George Allen & Unwin Ltd, London, reprinted in Routledge Classics (2009).

Russo, F. (2009). *Causality and Causal Modelling in the Social Sciences.* Springer.

Sack, A. T. (2006). Transcranial magnetic stimulation, causal structure-function mapping and networks of functional relevance. *Curr. Opin. Neurobiol*. 16, 593–599

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Sandis, C. (2012). *The things we do and why we do them.* Palgrave Macmillan.

Sartorio, C. (2009). Omissions and Causalism. *Nous*, 43 (3), 513-530.

Sartorio, C. (2016). *Causation and Free Will.* Oxford University Press.

Schaffer, J. (2004). Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation, in C. Hitchcock (eds.) *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell, 197-216

Schurger, A., Sitt, J.D., Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences of the United States of America* 109 (42), 2904-2913.

Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P.U., Wheatley, T. (2013). Barking up the wrong free: readiness potentials reflect processes independent of conscious will. *Exp. Brain Res.* 229 (3), 329-335.

Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P.U., Wheatley, T. (2015). Hypnotizing Libet: Readiness potentials with non-conscious volition. *Conscious Cogn.* 33, 196-203.

Schlosser, M.E. (2014). The neuroscientific study of free will: A diagnosis of the controversy. *Synthese*, 191 (2):245-262.

Shapiro, L., Sober, E. (2007). Epiphenomenalism. The do's and don't's. In: Wolaters, G., Marchamer, P. (eds.) *Thinking about causes: From Greek philosophy to modern physics.* University of Pittsburg Press.

Shea, N. (2013). Neural mechanisms of decision-making and the personal level. In Fulford, K.W.M., Davies, M., Graham, G., Sadler, J., Stanghellini G., Thornton, T. (eds.) *Oxford Handbook of Philosophy and Psychiatry.* Oxford University Press.

Searle, J.R. (1997). *The Mystery of Consciousness.* The New York Review of Books.

Searle, J.R. (2004). *Mind: A Brief Introduction.* Oxford University Press.

Searle, J.R. (2002a). *Consciousness and Language.* Cambridge University Press.

Searle, J.R. (2002b). Why I Am Not a Property Dualist. *Journal of Consciousness Studies* (12), 57-64.

Searle, J.R. (2010). Consciousness and the problem of free will. In : Baumeister, R.F., Mele, A., Vohs, K.D. (eds.) *Free wil land consciousness. How might they work.* Oxford University Press, 121-134.

Searle, J.R. (2015). *Seeing Things As They Are: A Theory Of Perception*. Oxford University Press .

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Labs Technical Journal* 27 (3), 379-423.

Shapiro, L. (2004). *The Mind Incarnate.* The MIT Press.

Shibasaki, H., Hallet, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology,* 117, 2341- 2356.

Shoemaker, S. (2001). Realization and mental causation. In G. Gillett & B. Loewer (Eds.), *Physicalism and its discontents.* Cambridge University Press, 74–98

Shoemaker, S. (2003). Realization, micro-realization, and coincidence. *Philosophy and Phenomenological Research,* 67(1), 1–23.

Shoemaker, S. (2007). *Physical Realization.* Oxford University Press.

Shoemaker, S. (2013). Physical realization without preemption. In Gibb, S.C., Ingthorsson, R. (eds.) *Mental Causation and Ontology*. Oxford University Press.

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

Sinnott-Armstrong, W., Nadel, L. (2011). *Cosncious Will and Responsibility. A Tribute to Benjamin Libet.* Oxford University Press.

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., Haggard, P. (2004) Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience,* 7 (1), 80-84.

Soon, Ch.S., Brass, M., Heinze, H-J., Haynes, J-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience,*

Sosa, E., Tooley, M. (1993). Introduction. In Sosa, E., Tooley, M. (eds.) *Causality*. Oxford University Press.

Strawson, G. (1986). *Freedom and Belief.* Oxford: Clarendon Press.

Strawson, P.(1962). Freedom and resentment. *Proceeddings of the British Academy*, 48, 187-211

Stein, N. (2012). Causal Necessity in Aristotle. *British Journal for the History of Philosophy*, 20(5), 855–879.

Spinoza, B. *Ethics* available at: https://www.gutenberg.org/files/3800/3800-h/3800-h.htm

Taylor, Ch., Dennett, D. (2011). Who's *Still* Afraid of Determinism? Rethinking Causes and Possibilities. In Kane, R. (eds.) *The Oxford Handbook of Free Will.* Second Edition. Oxford University Press. 221-240.

Timpe, K.(2013). *Free Will. Sourcehood and its Alternatives.* Bloomsbury. .

Trevena, J.A., Miller, J.(2002). Cortical movement preparation before and after a conscious decision to move. *Conscious Cognition,* 11 (2), 162-90, discussion 314-25.

Trevena, J., Miller, J. (2010). Brain preparation before a voluntary action: evidence against unconscious moment initiation. *Conscious Cogn.* 119(1), 447-456.

Tse, P.U. (2013). *The Neural Basis of Free Will: Criterial Causation.* The MIT Press.

Van Fraassen, B. (1980). *The Scientific Image.* Oxford University Press.

Van Inwagen, P. (1983). *An Essay on Free Will.* Oxford University Press.

Van Inwagen, P. (2013). A Modal Argument for Incompatibilism. In Russel, P., Deery, O. (eds.) *The Philosophy of Free Will. Essential Readings from the Contemporary Debates.* Oxford University Press, pp. 117-125.

Vicari, G. (2008). *Beyond Conceptual Dualism. Ontology of Consciousness, Mental Causation and Holism in John R. Searle's Philosophy of Mind.* Brill.

Von Wright, G.H. (1971). *Explanation and Understanding.* Cornell University Press.

Vargas, M. (2010). Responsibility in a world of causes. *Philosophical Exchange*, 40, 56-78.

Vihvelin, K. (2013). *Causes, Laws, and Free Will Why Determinism Doesn't Matter.* Oxford University Press.

Vohs, K.D., Schooler, J.W. (2008) The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science*, 19(1), 49-54.

Walter, S. (2008). The supervenience argument, overdetermination, and causal drainage. *Philosophical Psychology,* 21, 673–696.

Watson, D., Clark, L.A., Tellegen, A. (1988) Development and validation of brief measures of positive and negative affect. The PANAS scales. *Journal of Personality and Social Psychology,* 54, 1063-1070.

Wegner, D.M., Wheatley, T. (1999). Apparent mental causation. Sources of the experience of will. *Am. Psychol.* 54 (7), 480-92.

Wegner, D.M. (2002) *The Illusion of Conscious Will*. Cambridge Mass. MIT Press

Weinberg, R.A. (1985). The Molecules of Life. *Scientific American* 253 (4), 48-57.

Wiggins, D. (1973). Towards a Reasonable Libertarianism, In: Honderich, T. (eds), *Essays on Freedom of Action*, London: Routledge & Kegan Paul, 33–61

Wilson, G., Shpall, S. (2016) Action. *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/action/>.

Wolpert, D.M., Ghahramani, Z., Jordan, M.(1995). An Internal Model for Sensorimotor Integration. *Science,* 269 (5232), 1880-1882.

Woodward, J. (2003). *Making Things Happen.* Oxford University Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In: Hohwy, J., Kallestrup, J. (eds.) *Being reduced: New essays on reductive explanation and special science causation.* Oxford University Press, 218-262.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research* 91 (2), 303-347.

Yablo, S. (1992). Mental causation. *Philosophical Review,* 101 (2), 245-280.

Zimmerman, M. (1988). *An Essay on Moral Responsibility.* Rowan &Littlefield.