Minds, Brains, and Desert

ACTA PHILOSOPHICA GOTHOBURGENSIA 35

Minds, Brains, and Desert

On the relevance of neuroscience for retributive punishment

Alva Stråge

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 978-91-7346-530-4 (print)
ISBN 978-91-7346-531-1 (digital)
ISSN 0283-2380

The publication is also available in full text at:
http://hdl.handle.net/2077/60338

Academic thesis in Theoretical Philosophy
Department of Philosophy, Linguistics, and Theory of Science
University of Gothenburg

Distribution:
Acta Universitatis Gothoburgensis
PO Box 222
SE-405 30 Göteborg, Sweden
acta@ub.gu.se

Cover design by Alva Stråge
Print: BrandFactory, Gothenburg, 2019

For Alice & Ellen

Abstract

Title: Minds, Brains, and Desert
Author: Alva Stråge
Language: English
ISBN: 978-91-7346-530-4 (print)
ISBN: 978-91-7346-531-1 (digital)
ISSN: 0283-2380

Keywords: Desert, Responsibility, Philosophy of Mind, Neurolaw, Retributivism, Folk Psychology, Folk Morality

It is a common idea, and an element in many legal systems, that people can deserve punishment when they commit criminal (or immoral) actions. A standard philosophical objection to this retributivist idea about punishment is that if human choices and actions are determined by previous events and the laws of nature, then we are not free in the sense required to be morally responsible for our actions, and therefore cannot deserve blame or punishment. It has recently been suggested that this argument can be backed up by neuroscience, since neuroscientific explanations of human behavior leave no room for non-determined free actions. In this thesis, an argument of this sort is discussed. According to this argument, that I call "the Revision Argument", we should revise the legal system so that any retributivist justification of punishment is removed. I examine some objections to the Revision Argument according to which compatibilism about free will and responsibility is a morally acceptable basis of retributive punishment. I argue that these objections have difficulties in providing a plausible account of the relevant difference between people who deserve punishment for their actions and people who do not. Therefore, I argue that they fail to refute the conclusion of the Revision Argument.

Acknowledgments

First of all, I wish to thank Susanna Radovic, my main supervisor, and Gerben Meynen, my second supervisor, for their efforts, especially during the last year of writing. You have had hard times getting me on track (with various success) and you have had great patience with my stubbornness during this process. Thank you so much for sharing your expertise and time with me. Peter Andiné, my third supervisor, has provided me with valuable insights in fundamental questions of forensic psychiatry. Thank you.

I also want to thank Björn Haglund and Helge Malmgren who were my supervisors as an undergraduate. They were always interested and supportive when I came to them with my drafts and ideas, despite my not so impressive philosophical skills. Their friendly attitude made philosophy fun and exciting, and made me start thinking about going for a PhD, instead of leaving philosophy and start doing something "real."

Thanks to Karl De Fine Licht and Jakob Winther-Forsbäck for help and encouragement when I decided to apply for a PhD position.

I am immensely grateful for all my wonderful colleagues and friends at the department of Philosophy, Linguistics and Theory of Science at the University of Gothenburg. A special thanks to Thomas Hartvigsson and Ida Hallgren, who have been my roomies for the last years. Tomas who always shares refreshing stories from his life outside academia, and Ida who teaches me the art of negative thinking. Ellen Breitholtz has been a friendly face since I started my PhD-studies, always encouraging and helpful.

The participants of the Gothenburg research seminar in theoretical philosophy have provided me with many helpful comments and suggestions during the years. Thanks to Robert Hartman, Martin Kaså, Felix Larsson, Anna-Sofia Maurin, Per Milam, Susanna Salmijärvi, Marco Tiozzo, Anders Tolland, Maximilian Zachrau, and everyone else who has attended. A special thanks to Stellan Petterson who carefully read and commented on a draft of the first two chapters of the thesis, and to Ylwa Sjölin-Wirling who gave me valuable suggestions of how to phrase the Revision Argument, and besides

that has been my involuntary source of information about various formal matters in the final parts of thesis-writing. I am also grateful to Sofia Jeppsson who agreed to read a full draft of the thesis, and provided many valuable comments and suggestions.

Lena Eriksson – thank you so much for your firm guiding skills when things did not look so promising from my point of view. It really gave me the energy to keep going. I also owe a special thanks to Frans Svensson for reading parts of the manuscript and discussing it with me, providing me with valuable suggestions, and for supporting me in a stressful period of thesis writing.

Thanks to John Eriksson for being such a supportive and good friend. The lunch room would not be the same without you. All the lunch room people on the third floor deserve a special thanks for their spiritual discussions.

Helena Bjärnlind has always been helpful and kind when I drop by with big and small administrative questions over the years. Thank you.

I have been able to go to a number of conferences and discuss my work thanks to generous funding from and Stiftelsen Oscar Ekmans Stipendiefond, Kungliga and Hvitfeldska Stiftelsen, Adlebertska Stipendiestiftelsen, and Kungliga Vitterhets Historie och Antik Akademien. Kungliga and Hvitfeldska Stiftelsen, Adlerbertska Stipendiestiftelsen and Stiftelsen Petersenska hemmet have also provided extra funding which made it possible to extend my studies and finish the thesis. I am very grateful for this.

My parents have been extremely supportive and helpful during these years, and especially during the last one when it was difficult to combine intense writing with kids. Thank you so much for all your help in various matters. Also, thank you both for raising me in the spirit that it is perfectly alright to start doing things that one has no idea about how to finish, and for never show any doubt about that I am capable of doing whatever I set myself to do. Thank you, Mum, for telling me that I was born goal-oriented so many times that I started to believe it myself. That belief has been very much needed in this process. Also, a special thanks to my sister Hillevi who has provided me with some spiritual guidance and different kinds of support at various stages of the writing process.

Peter – thank you for being such a great co-parent, and for all the support during these years. Thanks also to Ingrid and Berno who are wonderful grandparents to the kids, and always kind and helpful.

To my beloved friends: thank you for still being there even though I have been a useless company for way too long. Marie, I cannot think of a more loyal, kind and loving friend than you. Thank you for not giving up on me! Emma & Lisa – always supportive, always inspiring, and always up for a night out – you give me hope about life after this thesis. Jenny, thank you for your ability to discuss important and less important things in detail, and for being the best roomie ever during my first years as a philosophy-student. Thank you all for letting me be part of the bok-klubb although I never ever read a single word.

Thanks to Sara for sharing some life-wisdom of yours during a stressful period of life. It was very much needed at the time, and helped me from being too disoriented.

Ragnar, I honestly could not have done this without you. Partly because of our endless philosophical discussions (that do not always end well) but most of all because you are much more encouraging, kind, patient and loving than I could ever have wished for, from anyone.

Two persons have been my anchors during these years. Regardless of the ups and downs in academic achievements (and life in general), my kids, Alice and Ellen, continuously remind me about what I really care about at the end of the day. Thank you for your patience when I have worked day and night, for helping me count how many pages there are left to write, for not complaining about my absent-mindedness, for your inspiring ideas about all the fun things we will do when the thesis is finished. I love you more than there are stars in the universe(s). I dedicate this book to you.

Contents

1 Introduction

1.1 Background and outline of the thesis

According to retributive theories of punishment, someone who commits a wrongful act can morally deserve to be punished. Retributivism is, arguably, part of common sense morality, in the sense that people in general think that someone who commits a wrongful act may deserve to be punished. This moral judgment is, at least partly, based on a particular common sense understanding of human behavior: that people can choose how to act and are therefore responsible for their actions. In this thesis, I will refer to common sense morality as "folk morality", and common sense understanding of behavior as "folk psychology." Retributivism is, moreover, also an element in many legal systems, in the sense that at least part of the moral justification for punishing people who commit crimes is that they deserve it.

In this thesis, the discussion will proceed from an argument that I call "The Revision Argument." According to this argument, the folk moral judgment that people can deserve punishment when they commit wrongful actions is built upon the folk psychological belief that people in such cases can act freely, since people in general have a libertarian free will. But, according to the Revision Argument, we have evidence pointing in the direction that there is no libertarian free will. This means that the folk moral judgment that people can deserve punishment lacks justification (since according to folk morality, people deserve punishment only if they have acted out of libertarian free will, but the belief that people can act out of libertarian free will turns out to be false.) Since the legal system, according to the Revision Argument, gains legitimacy from folk psychology and folk morality in the sense that people in general must, to a sufficient degree, support the way the legal system works, including its reasons for punishment, legal retributive punishment turns out to lack moral justification.

Hence, the legal system should be revised in the sense that we should remove retributive elements from its punishment practices.

The Revision Argument is a version of an argument originally put forth by Joshua Greene & Jonathan Cohen in their seminal article "For the Law, Neuroscience Changes Nothing and Everything" (2004). Greene & Cohen argue that neuroscience provides an explanatory framework for human thinking and action that makes us realize that many of our folk psychological assumptions of why people act in certain ways are false. For example, the assumption that people can act freely, in the sense of choosing how to act on the basis of their libertarian free will, has no support in a neuroscientific explanation of actions, since the brain, regardless of its complexity, works strictly in accordance with input-output principles. Greene & Cohen argue, further, that if there is no libertarian free will involved when people act, people do not deserve punishment in the way required according to retributivism. And then we should not punish people according to retributive principles.

Greene & Cohens's argument has gained considerable attention, especially within an interdisciplinary research field called "neurolaw" in which the implications of neuroscience to the law and legal practices are discussed (see e.g., Meynen, 2018 for an overview.) The main aim in this thesis is to scrutinize three major objections to the claim that we have reasons to revise our legal punishment practices. These objections all defend the view that compatibilism about free will and/or responsibility can provide what is required for desert-attribution and, consequently, justify legal retributive punishment.

The main focus in the discussions to follow is whether compatibilism is able to account for what will be referred to as "basic desert", which, according to many philosophers (and I agree) is a necessary condition for legal retributive punishment to be morally justified. I will explain what basic desert is in chapter two. The arguments considered in this thesis are all such that they are supposed to be compatible with the metaphysical doctrines of physicalism and determinism. Basic desert must, hence, be compatible with determinism and physicalism. I will describe how "physicalism" and "determinism" are interpreted in sections 1.3 and 1.4. In addition to these metaphysical constraints, I argue that the justification of legal

retributive punishment also must meet the demands of what will be referred to as "the Principle of Relevant Difference." This principle is related to what R.M. Hare (1952) calls the "ethical supervenience thesis", according to which there can be no moral difference between two states of affairs, events, actions or agents without there being some natural difference between them. I will suggest that this thesis can be developed in light of the plausible view that it is not enough to have *a* natural difference: *any* natural difference cannot do the job with regard to a specific moral difference – it has to be a difference that is intuitively of *a relevant kind*. However, it is not a straightforward matter to determine what is a relevant difference and what is not, and different natural properties may be relevant in relation to different moral properties. One of the premises in the Revision Argument states that libertarian free will is the relevant difference with regard to basic desert: possession of libertarian free will is, according to the Revision Argument, the *relevant difference* between two people who commit similar wrongful acts, but where only one of them deserves punishment. Thus, if none of them has libertarian free will, none of them deserves punishment. However, according to the objections to the Revision Argument that are to be discussed, libertarian free will is *not* the relevant difference between someone who deserves punishment and someone who does not. Instead, according to these objections, the relevant difference between someone who deserves retributive punishment and someone who does not has to do with certain mental capacities to identify and react on reasons, and these capacities are fully compatible with determinism.

I will argue that in order to successfully defend such a compatibilist claim one must be able to pick out a natural property that is intuitively relevant for basic desert, and which is sufficiently different from other natural properties, properties that are not base properties of basic desert.[1] This analysis will play a central role in my argument. Relying

[1] That there must be a relevant difference between cases in which we ascribe moral responsibility and cases in which we do not is also argued for by Derk Pereboom (2001, 2002), although he focuses on (the absence of) a relevant difference between manipulation cases in which we, intuitively, do not want to hold someone morally responsible, and "ordinary" cases in which we, intuitively, want to hold the agent morally responsible.

on the Principle of Relevant Difference, based on Hare, I will argue that if compatibilist theories cannot provide such an intuitively relevant difference between the base properties of basic desert and other natural properties that are not base properties for basic desert, then desert-attribution violates what Jaegwon Kim calls the "consistency requirement", which is based on the well-established moral principle known as "the principle of universalizability." This principle, in turn, relies on the intuition that ethical judgments should be generalizable in some sense, i.e., the idea that that like cases should be treated alike (Kim, 1984, pp. 161-162).

It is worth pointing out that while the Principle of Relevant Difference is a normative principle, this thesis is not primarily intended to be a contribution to the normative discussion of what is required for desert-attribution. Rather, the Principle of Relevance Difference is considered to be part of folk morality, and folk morality is, in turn, the enterprise that legitimizes the legal system. If there are features in the legal system that violate fundamental folk moral intuitions – for example, if the legal system attributes basic desert to people in a way that violates the consistency requirement – the legitimacy of legal responsibility attribution is jeopardized. Legal practices must be legitimate, or they need to be revised (as a consequence of folk morality.) To be clear, even though legal responses to wrongdoing are the topic of this thesis, the thesis itself is philosophical in nature. Its basic arguments are general arguments that are not derived from one particular legal system. In fact, the arguments that are discussed are fundamental theoretical arguments that have been developed in the neuro-legal and philosophical debate about the implications of neuroscience – and other allegedly deterministic branches of science – for the law and legal practices.[2]

The structure of the book is as follows: in chapter one, *Introduction*, I will provide a brief overview of some philosophical discussions that are in focus in this thesis. In chapter two, *The Revision Argument*, I introduce the Revision Argument. After outlining the argument, I will

[2] This approach is in line with the analysis by many other authors, for instance, it can be found in the work of Pardo & Patterson (2010, 2013), Vincent (2013), Moore (1997, 2016), Morse (2004, 2009), Greely (2011), Sapolsky (2006).

discuss some different notions of responsibility in order to analyze what responsibility means in the Revision Argument. In chapter three, *First Objection: Legal Retributive Punishment Does Not Require Free Will,* I will discuss an objection put forward by Stephen Morse, according to which the Revision Argument is flawed because (1) libertarian free will is irrelevant for legal responsibility, and (2) libertarian free will is also not a requirement for responsibility and retributive punishment according to folk psychology and folk morality, and (3) to the extent that people do talk about free will as a requirement for responsibility and retributive punishment, free will compatibilism provides a theoretical framework that can meet these requirements.

I will show that there is a substantial disagreement among both legal scholars and philosophers about the first claim. Regarding the second claim, I turn to some experimental philosophy studies concerning people's intuitions about the compatibility of determinism, free will, and responsibility. These studies do not seem to provide any straightforward answers about people's intuitions of these things, besides the fact that the responses seem to be sensitive to the set-up of the experimental case. The third claim is analyzed in chapter four, *Is Compatibilism a Secure Basis of Retributive Punishment?* in which the question is discussed in light of some central compatibilist ideas about sufficient conditions for basic desert, mostly centered around reasons-responsiveness.

From the discussions in chapter three and four, I will argue that this first objection fails. The compatibilist theories considered cannot meet the demands provided by the Principle of Relevant Difference in combination with the metaphysical constraints of physicalism and determinism.

In chapter five, *Second Objection: Conceptual Confusion about the Nature of Mental States*, I will discuss an objection put forward by Michael Pardo & Dennis Patterson against the Revision Argument but also against to the argument I provided in chapter four. Pardo & Patterson argue that the Revision Argument, but also my argument to the effect that Morse's objection fails to refute the Revision Argument, are based on a conceptually confused view of mental states. If their argument is sound, it may still be the case that free will compatibilism

can provide what is required for basic desert. However, I will argue that this objection fails, too. Even though I agree that a plausible account of mental states allows for multiple realizability and, hence, rule out the possibility of reducing them to brain states in a straightforward manner, mental causation cannot plausibly be disconnected from brain processes. And since mental causation plays a central part of basic desert, I will conclude that the second objection, too, fails to meet the requirements of the Principle of Relevant Difference in combination with the metaphysical constraints of physicalism and determinism.

In chapter six, *Third Objection: The Limited Relevance of Neuroscience & Philosophy for Folk Psychology, Folk Morality & the Law,* I will discuss the worry that regardless of the theoretical relevance of the discussions in chapters three to five, legal practice is ultimately legitimized by folk psychology and folk morality. For different reasons, this point would render the law and legal practices "immune" to the theoretical concerns regarding retribution raised in this thesis. I will, however, argue that the "folk" understanding of the world is not immune to such theoretical concerns, and that this fact implies that the sciences are also of relevance for the law – in a substantial sense.

Moreover, I will offer an alternative description of folk psychology and folk morality, according to which it is plausible that folk psychology and folk morality implicitly support the Principle of Relevant Difference as it has been discussed in the previous chapters. If this account is accurate, and my conclusion that the arguments provided by Morse, Pardo & Patterson fail to account for basic desert, the Revision Argument stands: the retributive element in the current legal system lacks justification and thus, these parts needs to be revised.

In chapter seven, *Summary & Concluding Remarks,* I provide a summary of the discussions in chapters two to five, and move on to briefly discuss some implications of the conclusions in these chapters, as for example, what the consequences may look like if we revise the legal system in accordance with the conclusion of the Revision Argument.

The structure of the remaining parts of this introduction is as follows: section 1.2 will provide a brief overview of the neurolaw

debate, in order to illustrate the context in which the Revision Argument is developed. In section 1.3 and section 1.4, I will describe how "physicalism" and "determinism" are interpreted in this thesis. Section 1.5 will be concerned with an introduction to the difference between compatibilism and incompatibilism about free will, a difference that will play a central role throughout the book. Section 1.6 will be briefly concerned with some philosophical theories of mental states. This topic is not explicitly addressed in the Revision Argument but is of importance to the discussions to come, especially in chapter five. The chapter ends with a short summary in section 1.7.

1.2 Neurolaw

Neurolaw is an interdisciplinary field that links neuroscientific research findings to the law and legal practices. Neuroscience has developed quickly in the past decades and it has received much attention from scientists from other branches, as well as from the political sphere. There are a number of politically initiated projects that receive immense funding. For example, the European Union's Human Brain Project is a ten year project (it started in 2013) and aims to build a research infrastructure to help advance neuroscience such as brain simulation and neuroinformatics (i.e., access to shared brain data), medicine (e.g., access to patient data, identification of disease signatures) and computing (e.g., development of brain-inspired computing, (e.g., use of robots to test brain simulation) but also ethical and societal implications of the technical developments in these areas ("Human Brain Project", 2017). A project called "the Brain initiative" was announced by the Obama administration in the U.S. in 2013, with the goal of supporting the development of innovative neuroscientific techniques in order to deepen the understanding of the of the human mind and to improve how to treat, prevent, and cure disorders of the brain (The Brain Initiative, 2019).

Neuroscience is regularly appearing in the courtroom, and the fast development of different techniques that enable more detailed information about the correlations between brain functions and behavior have made neuroscience increasingly used in order to, for

example, provide data, supposedly relevant for the question of legal responsibility.[3]

The Anglo-American adjudicatory system has long attempted to include the special knowledge of skilled witnesses and scientific experts in the area of brain science (Hall, 2004). Since 2005, the use of neuroscientific evidence in American courts has been tripled, and such evidence is used in 25% of all murder-cases. The situation in the U.S. is not unique. Research in England and Wales, Canada, the Netherlands and Singapore shows that defense lawyers in all jurisdictions make use of neuroscientific evidence to defend their clients. Typically, if neuroscientific techniques are introduced in the courtroom, it is often in cases in which the charges are serious and severe punishment is possible. For example, neuroscientific evidence has been critical in overturning convictions for murder and reducing convictions from murder to manslaughter (Catley, 2016).

It can be argued that a legal system that incorporates scientific facts into its practices is more reliable compared to a system that does not. Think of lies: if we could depict what happens in the brain when someone is lying (compare to when she is sincere, or mistaken) it would be easy to find out who is lying and who is not by making use of a brain scan.[4]

Meanwhile, it can also be pointed out that (neuro)science often fails to provide the clear answers that are sought after in a legal context. For example, even though a brain scan may reveal certain facts about someone's brain functions, it is not clear to what extent such facts can provide any additional explanatory power beyond what we already know from behavioral evidence such as descriptions of the defendant's actions, or first or third person reports about her mental states. One highly influential voice who cautions against the use of neuroscience in the legal context is the legal scholar Stephen Morse. He maintains that those who put too much faith in the explanatory power of neuroscience suffer from "the brain overclaim syndrome" (Morse, 2006). Neuroscience cannot, at least not in the vast majority

[3] See e.g., Catley & Claydon (2015), Farahany (2015), Chandler (2015) and de Kogel & Westgeest (2015).

[4] See e.g., Farah et al (2014) for discussion.

of cases, deliver what we need to answer the relevant legal questions, he argues. One reason for this shortcoming, Morse points out, is methodological: the neuroscientific studies at hand are often made on a small number of people, why we must be careful drawing general conclusions from them. And even if we can draw certain conclusions from such studies at group level, this data still does not provide us with conclusive evidence about specific individuals. For example, even if we can identify certain correlations between EEG patterns and e.g. levels of consciousness on a group level, these correlations does not necessarily maintain in the individual case (Morse, 2006, pp. 403-405). However, Morse acknowledges that there are some contexts in which neuroscience can contribute to legal controversies, such as, for example, in cases when the defendant has received a blow to the head, and it is unclear if he acted consciously (Morse, 2006, p. 401).

Hence, on the one hand, it can be argued that neuroscientific evidence is a potentially valuable resource for the legal system, and the system should, therefore, use neuroscientific evidence in order to improve its methods – e.g., for identifying defendants that deserve to be punished and defendants that should be excused. Or, as it is argued in the Revision Argument, to find out that some of our beliefs about human behavior are false. On the other hand, it can be argued that, at present, neuroscientific evidence is too unreliable to be used in court cases, and even that neuroscience has, in general, a very limited relevance for the understanding of human action in the legal context.

Neurolaw is not a homogenous research field, and the possibilities and problems that are associated with questions of how neuroscience can be useful for the legal system vary within different areas of neurolaw. In the next section I will briefly describe some different areas of neurolaw, and specify what will be focused on this thesis.

1.2.1 Different areas in the neurolaw debate

The neurolaw debate is concerned with a wide array of questions, stretching from technical issues regarding the practical use of neuroscientific data in the courtroom, to moral and legal concerns about the foundations of our legal system – the latter being the topic

of this thesis. Adrian Petoft (2015) provides a useful two-part distinction between practical and theoretical neurolaw discussions. Practical discussions are, in Petoft's view, those that focus on civil and criminal responsibility in legal processes, such as how neuroscientific data is used in the court room and the implications of this use. Theoretical research focuses on how neuroscientific research contributes to our general understanding and knowledge of the mind, and how it applies to philosophical questions concerning mental states and free will which are relevant for the question of criminal responsibility.

Gerben Meynen (2016) distinguishes between three different areas within the neurolaw debate: Assessment, Intervention and Revision. These areas are described briefly below.

Assessment

In criminal law, lawyers have to answer a wide range of questions. Common examples are: does the defendant have a mental disorder? Is she competent to stand trial, is she legally insane? What is the risk for recidivism for this particular offender? What does the witness remember exactly – is she lying? Is the prospective juror biased against certain groups of people? (Meynen, 2016, p. 3). Part of neurolaw research focuses on how neuroscientific techniques can assist in answering such questions which concern the evaluation or assessment of an individual. In the task of assessing the presence of mental disorders, neuroscientific techniques such as for example brain scans could contribute with valuable information about e.g., deviant brain structures that may have implications for a person's cognitive capacities (Meynen, 2016, pp. 134-138).

A worry often mentioned in the assessment domain, which was mentioned above, is the problem of generalization. Even if we find statistically significant correlations between certain brain functions and certain mental states on a group level, conclusions concerning a particular individual cannot easily be drawn (see e.g., Pardo & Patterson 2013, p. 145 and Morse, 2006, pp. 403-406). However, one could argue that the assessment of e.g. mental health and the degree of control over one's actions have reliability-limitations no matter what methods are being used. When a defendant's mental capacity is

evaluated through forensic psychiatric or psychological examination, these methods are also based on empirically based tests, and the validations of such tests are also ultimately based on general statistics that are applied to individual cases.

Besides the potential usefulness of assessing legally relevant mental health problems, neuroscientific methods have been used in order to develop techniques for lie detection.[5] The legal utility for a reliable lie detector is obvious, but this area of research is perhaps suffused by even more controversy than the area of mental health assessment. Questions of e.g., under what conditions these techniques should be used and the reliability of lie detection techniques raise fundamental issues of how neuroscientific techniques can be relevant in legal contexts. For example, it is hard to develop experimental settings that are sufficiently similar to the contexts in which lie detection would be most helpful (see Pardo & Patterson, 2013). For example, it is plausible to think that the difference between what is at stake in the cases of lying in order to avoid a prison sentence and lying in an experimental setting, is reflected at the neural, cognitive and emotional level as well (Pardo & Patterson, 2013, p. 109).[6]

A third task for which we can see a potential use of neuroscience is to assess whether people are inclined to be affected by implicit bias in their judgements. This factor is relevant for jury selection and for assessing reliability in testimonies (Greely, 2011, p. 1225).

Intervention

The intervention domain covers questions about both practical and theoretical possibilities of using neuroscientific techniques in order to affect people's thinking and action, but also discussions of the ethical implication of using such interventions in a legal context.

Interventions can be of different sorts, and a common division between forms of interventions is treatment, enhancement and manipulation (Meynen, 2014, p. 820). *Treatment* concerns the question

[5] See e.g., Farah et al (2014).

[6] See Pardo & Patterson 2013, pp.79-120 for a detailed discussion of the difficulties for the prospect of brain states replacing behavior as the criteria for lies or deception. This will be further discussed in chapter five.

about current and future possibilities regarding treatments that can, for instance, help reduce the risk of recidivism. In this domain, the aim is ultimately to change brain functions in order to reduce criminal behavior. The *enhancement* area also involves discussions about possible consequences (ethical and otherwise) that different enhancement methods may have both within and outside of the legal sphere.

Discussions about *manipulation* are related to cases in which the behavior is controlled, in some sense, by external sources. For example, if a deep brain stimulation device is implanted in order to control obsessive-compulsive disorders, is the result of the workings of such a device within the agent's responsibility? If such a device turns out to have a negative effect on a person's behavior, but perhaps also on her preferences, beliefs, desires and so on, in a way that have legal consequences - who is responsible for her actions? Or, how should the law deal with a scenario in which a deep-brain stimulation device is hacked by someone who manages to manipulate the user of the device to develop certain preferences in order to make her perform criminal actions? In legal cases, manipulative circumstances could play a mitigating role, but usually the history of people's preferences are not exculpatory factors as such (Bublitz & Merkel, 2013, p. 340).

With regard to the domain of enhancement, if safe, reliable and effective techniques to enhance mental performance are developed, are there some groups that should be obliged to enhance their cognitive performance in order to maximize the likelihood of good outcomes, such as, for example, surgeons, or pilots? [7] If so, what are the consequences for responsibility attribution if someone who is obliged to take it refuses? Another aspect of the enhancement discussion is the question of whether cognitively enhanced people should be held responsible for their actions to a higher degree since mental capacities often are taken to co-vary with responsibility attribution. The idea behind this scenario is that if responsibility is diminished when mental capacities are reduced, and restored when the mental capacities are regained, then it seems as if we assess people

[7] C.f. Vincent, 2013, p. 326.

differently when it comes to responsibility, depending on what mental capacities they have. But then again, if the mental capacities are going beyond what is "normal" – should we expect this person to shoulder more responsibility than the "normal" person? This kind of argument is disputed and there are many counter-arguments against the plausibility of such views. For example, Nicole Vincent suggests that responsibility is a "threshold" concept, at least insofar as the law is concerned: the law imposes an objective rather than a subjective standard of care onto everyone, and as long as you reach that standard (or threshold) you cannot be blamed for if things goes wrong *even if* you would have the capacity to do more than just reach that minimum standard (Vincent, 2013a, p.188). Similar ethical considerations are figuring in the treatment domain. For example, to what extent should a defendant with addictive problems be coerced to follow treatment? (see e.g., Meynen, 2014, p. 821).

Revision

In the revision domain, research focuses on whether neuroscientific findings should lead to revisions in the law or legal practices. In this domain, Petoft's division between practical and theoretical questions becomes particularly relevant. A central part of this discussion revolves around the question of what neuroscientific findings means for our responsibility practices, both in law but also with regard to moral responsibility. Some argue that neuroscientific explanations of human behavior support the hypothesis that free will is an illusion, and that this lack of free will, in turn, comes in conflict with some fundamental elements in moral thinking as well as in our legal system. This hypothesis is based on the intuition that the moral basis of blaming and praising people is built on the assumption that they have a genuinely free choice. Contestants of this argument claims that this interpretation is a misunderstanding of what matters to our responsibility practices: they may argue that free will is compatible with neuroscientific explanations, since we have free will in virtue of possessing certain capacities and abilities, which is consistent with a naturalistic understanding of behavior.

This theoretical discussion concerns the theoretical underpinnings of the legal system, i.e., assumptions and beliefs that are not

necessarily are part of the written law and regulations but are fundamental in the sense that many aspects of the legal system make sense only in light of these assumptions. For example, if we did not assume that people can adapt their behavior to legal standards, then most parts of the legal system seem unmotivated (Morse, 2007).

This discussion can be compared to what Petoft calls "practical discussions", which focus on how procedures within the legal system can be revised in light of neuroscience. One such example was mentioned in the previous section, concerning how neuroscientific methods could be used in coerced treatment of offenders with addictive behavior. Another example of such a practical discussion is the one concerned with how to treat juvenile offenders. Drawing on research about how the brain develops, it can be argued that a young person cannot be expected to have the same cognitive capacities as an adult, and consequently, young people do not deserve the same kind of legal treatment as adults (see on this topic Meynen (2016 and Steinberg (2013)). Another example considers how a better understanding of for example addictive behavior can lead to revisions in how addicts are treated with regard to responsibility and punishment, or more knowledge about certain mental disorders might revise the way we blame and punish mentally disordered offenders.

In this thesis, the discussion is restricted to what according to Petoft's distinction belongs to the theoretical part of the neurolaw discussion, and more specifically, to the theoretical part of the revision domain. The point of departure will be what I will call the Revision Argument, in which the central claim is that neuroscience gives us reasons to revise some fundamental assumptions in our legal system. The discussion surrounding this argument is directly related to the traditional free will and responsibility discussion, in the sense that it focuses on the question of how to combine a physicalist understanding of human behavior with the view that people are responsible because they have a genuine free will. The aim in this thesis is not to defend any particular idea of what it takes to act freely. Rather, I want to scrutinize whether compatibilism can provide what is required for the practice of retributive punishment to be morally justified.

Free will compatibilism can be contrasted to free will incompatibilism. The disagreement between these two views concerns the possibility of having a genuinely free will in a deterministic world. Compatibilists argue that we have a genuinely free will also in a deterministic world, while the incompatibilist refutes this claim. Determinism and physicalism play central roles in the free will discussion, and will do so throughout this book. In the following sections, I will give a brief introduction to relevant metaphysical doctrines, and explain how I will interpret them in the discussions to come. Next, I provide a brief overview of some major positions in the free will discussion.

1.3 Physicalism

Physicalism is a central notion in the discussions to follow, but it is a complex notion surrounded by disagreements of what it amounts to. For example, what is the basic claim of physicalism? Is it that everything *is* physical, or is it that everything must fit into a physicalist description of the world? For our discussion, it will suffice to understand physicalism in a manner suggested by e.g., Frank Jackson:

> Physicalism […] claims that a complete account of what our world is like, its nature, (or, on some versions, a complete account of everything contingent about our world), can in principle be told in terms of a relatively small set of favored particles, properties, and relations, the 'physical' ones. (Jackson, 1998, p. 6)

According to this description of physicalism, it is not ruled out that there are no non-physical phenomena. A number of theorists, as e.g., David Lewis (1986), David Chalmers (1996) as well as Jackson himself (1998) have suggested that we need a contingent global supervenience thesis that says something of the actual world and various worlds that are similar to the actual world, in order to illustrate how physicalism restricts the variation of phenomena that supervene on the physical. Jackson's version of such a supervenience thesis goes as follows:

> Any world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world. (Jackson, 1998, p. 12)

This basically means that even though there are phenomena in the world that are not physical, these phenomena supervene on the physical properties in the world. Henceforth, if there is a world *exactly like ours* with respect to physical properties, then this world will be exactly like ours with respect to all the non-physical phenomena we admit in our ontology.

One problem for the supervenience view is, according to some theorists, that it allows for non-reducible, non-physical properties. Since the supervenience thesis only restricts the world in the way that non-physical properties necessarily co-vary with the physical properties, it does not restrict what non-physical properties there are or the nature of these properties. In other words, the supervenience view is consistent with property dualism.

J.J.C Smart describes the property dualism argument as concerned with the following issue:

> Suppose we identify the Morning Star with the Evening Star. Then there must be some properties which logically imply that of being the Morning Star, and quite distinct properties which entail that of being the Evening Star. (Smart, 1959, p. 148)

If we apply this line of reasoning to the mind-body identity theory Smart concludes that there must be some mental properties that are logically distinct from physical properties: if we identify mental states and physical states by different criteria, then theses states seem to have distinct properties. The reasoning about identities resembles of Saul Kripke's argument for dualism, according to which identities, if true, are *necessarily* true. Kripke argues, in relation to the mind-brain identity theory, that cases of minds without brains are possible, as well as brains without minds. Since minds can exist without brains, and vice versa, minds and brains are not identical (Kripke, 1972, 1980, pp. 153-154). Henceforth, there are non-reducible, mental properties in the world.[8]

Kim (2002), argues that if we accept that mental properties are irreducible to physical properties, the physicalist thesis will nonetheless be different from the substance dualist thesis in the sense

[8] Property-dualism about mental states will be described in some more detail in section 1.6.4.

that insofar as non-reductive properties are causally effective, their causal relevance is due to physical properties. This view of causality is also defended by David Papineau:

> Mental occurrences have physical effects. For example, Eric Bristow's desire to score thirty-two at darts causes him to hit double sixteen. But such physical effects are also attributable to physical causes. The trajectory of Eric Bristow's dart is also caused by the arrangement of neurons in his brain and his consequent bodily movements. So, unless we want to say that such physical effects are overdetermined by two separate causes, which we clearly don't, we need somehow to view the mental cause and the physical cause as the same cause. (Papineau, 1990, p. 66)

Papineau's reasoning is based on the idea of "causal closure" which he formulates as the thesis "all physical effects have sufficient physical causes" (Papineau, 1998, p. 375).

There are objections to the principle of causal closure: for example, E.J Lowe (2000) argues that various forms of naturalistic dualism are consistent with the strongest physical causal closure principles that can plausibly be advocated. However, I take the principle of physical closure to be sufficiently well-established – even if not uncontroversial – for being used as a basic assumption about causation within a physicalist framework in the discussions to come. Yet, that mental causation either is identical with physical causation or implies overdetermination is not an obvious consequence of the acceptance of the causal closure of the physical realm. I will return to questions of physicalism, supervenience, and mental causation in chapters three, four and five, where they will be connected to questions of how moral properties supervene on natural properties.

1.4 Determinism

Determinism is a perennial topic in philosophy, and it is no exaggeration to say that it has played an absolutely central role in the free will discussion. However, exactly how to understand what determinism is, and what theoretical implications it has, is far from clear. As John Earman expresses it:

> […]some take the message of determinism to be clear and straightforward while others find it hopelessly vague and obscure; some take determinism to be intimately tied to predictability while others profess to see no such bond; some take determinism to embody an *a priori* truth, others take it to express a falsehood, and still others take it to be lacking in truth value; some take determinism to undermine human freedom and dignity, others see no conflict, and yet others think that determinism is necessary for free will; and on and on. Here we have, the cynic will say, a philosophy topic *par excellence!* (Earman, 1986, p. 1)

As Earman nicely illustrates, determinism is the subject of immense disagreement among philosophers, both with regard to its content, but also with regard to its truth-value, and its implications for human freedom. Part of the problem is that the philosophical relevance of the deterministic thesis is closely connected to theories in physics. And even though philosophers may speculate about what physical determinism really means, we often (quite naturally) lack insight in the complex details of how different physical theories work, and what role determinism plays in these theories. Earman continues:

> Classical physics is supposed by philosophers to be a largely deterministic affair and to provide the paradigm examples of how determinism works. Relativity theory, in either its special or general form, is thought merely to update classical determinism by providing for Newtonian mechanisms relativistic counterparts that are no less and no more deterministic. And it is only with the advent of the quantum theory that a serious challenge to determinism is supposed to emerge; the challenge is simply not that quantum mechanics is *prima facie* non-deterministic but that "no hidden variable" theorems show that, under plausible constraints, no deterministic completion of the quantum theory is possible. This picture is badly out of focus. Newtonian physics, I will argue, is not a paradise for determinism; in fact, Newtonian worlds provide environments that are quite hostile to determinism [...] The special theory of relativity rescues determinism from the main threat it faces in Newtonian worlds, and in special relativistic worlds pure and clean examples of determinism, free of artificial props, can be constructed [...] The quantum theory, of course, poses challenges of its own; but the first and foremost challenge is not to the truth of the doctrine of determinism but to its meaning in quantum worlds where the ontology may be nothing like that presupposed in the Newtonian and relativistic formulations of the doctrine. (Earman, 1986, p. 2)

However, as Henric Walter (2001) points out, how to exactly understand the quantum world is far from clear even among the most renowned physicists: "It does not help much to appeal to quantum theory's founding father Niels Bohr, nor to genius Einstein, nor to the acclaimed physicist Penrose to defend the 'true' or 'really correct' interpretation of it" (2001, p. 25).

Given this picture, it might seem as if the use of the notion of determinism may cause more troubles than solutions in a philosophical discussion of free will. But if we accept that the meaning of determinism in physics is obscure, and that even the most renowned experts are not united in how to understand determinism in relation to different explanatory frameworks of the physical world (such as quantum mechanics, Newtonian physics, and relativistic formulations) we can still make use of determinism as a metaphysical doctrine, similar to other metaphysical doctrines philosophers use in order to build theoretical frameworks. For example, in ethics some people may postulate that there are moral facts, and in metaphysics, some may postulate that there is a reality independent of us. Given that determinism is not an empirically established thesis, I think it is plausible to view the use of determinism in philosophical discussions in a similar vein: as a metaphysical doctrine that may or may not be true. When determinism is referred to in the following discussions, it is such a metaphysical hypothesis that I have in mind. Still, that does not mean that using the term "determinism" is without complications, but that at least (some of) the issues related to physics theories can be circumvented in this way.

In philosophy, a common characterization of determinism states that every event is causally necessitated by antecedent events. It can be summarized as the thesis that the facts of the past, in conjunction with the laws of nature, entail every truth about the future (see, e.g., McKenna & Coates 2015, O'Connor, 2005). Another way to put it is like Randolph Clarke does:

> […] determinism is the thesis that our world is such that any possible world that has the exactly same laws of nature and that is exactly like our world at any one point in time is exactly like it in every point in time […] I shall take determinism to conjoin this claim with the thesis that for every even *E* (except those beginning at the very first time, if

> there is a first time), at every time *t* prior to the occurrence of *E* there is some event (or some plurality of events) that occurs at *t* that deterministically causes *E*. One event is taken to deterministically cause another just in case, in every possible world in which the actual laws of nature obtain and in which the first event occurs, it causes the second. (Clarke, 2003, p. 4)

This thesis means that for every person, the facts of the past, in conjunction with the laws of nature, entail every truth about the person's future acts. Does free will have any place in such a framework? At face value, the above description of determinism seems to imply that people's actions are completely governed by factors beyond their own control. However, what it means to have control, and hence what it takes to exercise free will in a deterministic world is a topic of discussion. According to compatibilists, there is a sense of control that can be maintained in a deterministic world, and in this sense, we can have free will. The incompatibilist refutes this contention: according to her, free will is not compatible with us being completely determined. A brief description of these two positions is provided below.

1.5 Free will

In B.F Skinner's (1948) utopian novel *Walden Two,* the citizens live rich lives. They pursue arts, sciences, crafts and music. They enjoy what seems to be a pleasant existence with plenty of leisure. In a way, Walden Two is the freest place on earth, since the people living there have maximal freedom of choice and action. They can do anything they want to do. There is no coercion, and no punishment. No one has to be forced to do anything against his or her will. However, behavioral engineers covertly control people's wishes and desires. They can do anything they want since they have been conditioned not to want anything they cannot have.

Are the people in Walden Two free? In the novel, a philosopher visiting Walden Two argues that they are not, since all they have is *surface* freedom, whereas real freedom must also consist of *deep* freedom of will. Frazier, the fictional founder of the society, answers that there is no real loss. Frazier thinks there is no such deep freedom

of will: it is an illusion in the first place. We do not, and cannot, have deep freedom of will, neither inside nor outside Walden Two.

Skinner's novel illustrates the essence of the free will discussion. What is it to exercise one's free will? Do we have free will even if our wishes and desires are determined by factors beyond our control? The possible determining factors may include fate, God, the laws of nature, heredity, psychological or social conditioning, hidden controllers, and so on. But they all lead to the question whether we really are free. Many people appreciate that there is at least an apparent conflict between free will and determinism, but there is an immense disagreement whether they are truly incompatible, or that the apparent conflict disappears if we do some philosophical footwork. Many philosophers and scientists have argued that despite appearances to the contrary, determinism poses no real threat to free will, at least not the relevant kind of freedom or free will. This view that determinism is not a threat to the relevant kind of freedom or free will, is called "free will compatibilism". The opposite view that free will is not possible in a determined world is called "free will incompatibilism."

1.5.1 Compatibilism

Compatibilism about free will seeks to explicate an account of free will that is not threatened by the possibility that all actions are causally determined. Michael McKenna (2015) suggest that compatibilism's place in contemporary philosophy can be understood as a development with at least three stages. The first stage entails classical compatibilist theories with roots in historical writing by for example Thomas Hobbes in the 16th century, and David Hume in the 18th century. In the modern discussion such classical compatibilism was defended by e.g. A.J Ayer (1954) and J.J.C Smart (1961). The core in classical compatibilist theory is that freedom of the will only requires the absence of compulsion and coercion. When an act is caused by or carried out in line with an agent's desires and/or wills, then it is an act out of free will. The second stage departures the classical compatibilist view in the 1960's: one incompatibilist argument put forward by Carl Ginet (1966) to the effect that if one has not the

possibility to do otherwise, one has no free will (Ginet's argument is similar to what is known as "the consequence argument" put forward by van Inwagen, 1983); Harry Frankfurt's thought experiment against the intuition that responsibility presupposes the possibility of doing otherwise (Frankfurt, 1969); and P.F Strawson's descriptive account of responsibility practices and reactive attitudes in his seminal paper "Freedom and Resentment" (Strawson, 1962a). The third stage involves different kinds of contemporary forms of compatibilism that have been developed through the discussions and insights that characterized the second "transitional" stage, as for example, debates about the plausibility and implications of Frankfurt-style examples (see Kane, 2002, p. 17). I will discuss some compatibilist theories more thoroughly in chapter four.

1.5.2 Incompatibilism

Free will incompatibilism states that any free action must be an undetermined event: we can act freely only if determinism is false. It is worth pointing out that incompatibilism does not entail that there is no free will – it simply points out that free will and determinism are incompatible. One can be incompatibilist without taking a stance on the question about whether we have a free will or not – the incompatibilist claim is only that *if* determinism is true *then* there is no free will. Or, one can be an incompatibilist and take determinism to be true and, thereby, dismiss the possibility of free will. A third alternative is to be incompatibilist and embrace libertarianism. Libertarianism about free will is the position that people sometimes act freely in the sense that is incompatible with determinism: people sometimes act without being causally determined to do so. As was noted above, determinism, as it is interpreted in this thesis, is a metaphysical hypothesis, and there are those who argue that we have more evidence supporting the hypothesis that the world is not determined than we have for the hypothesis that it is. When accepting such a claim, libertarianism about free will seems more plausible than alternative theories of free will. However, libertarianism about free will is not a homogenous theory: there are at least three major categories of libertarians: event-causation libertarians, agent-

causation libertarians and non-causal libertarians (O'Connor & Franklin, 2019). I will provide a brief description of the basic ideas in each of these views below.

Event-causal libertarianism

Event-causal libertarians hold that some of a person's actions are self-determined and that self-determination requires nondeterministic causation.

One of the most prominent contemporary defenders of this type of libertarianism is Robert Kane (see e.g., 1985, 1989,1996, 2000). In Kane's libertarian theory, free will and moral responsibility have close connections, and he thinks that people can be morally responsible for actions that are causally determined by their character. But he insists that in order to be morally responsible, at least some of the prior actions that have contributed to that character have not been causally determined. Kane calls these actions "self-forming actions" – or SFAs – since it is, at least partly, because of these actions that our character is like it is (Palmer, 2104, p. 4). In one of Kane's examples of such a self-forming action, we are to imagine a business woman on her way to an important meeting when she comes across a person in need of help. The woman is torn by doing the moral thing (to stop and help) and doing the self-interested thing (carrying on to her meeting). She recognizes the reasons to do both and regard neither of the sets of reasons as weighing more than the other. Kane argues that in light of these conflicting motivations, the woman must make a mental effort to get her ends or purposes sorted out or to "set" her will in one way or another, that may initiate action in one way or another (Kane, 1996, p. 126). And it is the event of her making this effort, in combination with the event of her having reasons to perform it that non-deterministically cause her action (Palmer, 2104, p. 6). At his point, in his work, Robert Kane also refers to a situation of "chaos" and "quantum events" in the person's brain. "Non-determinism" in this context should be understood as that the reasons one has for performing an action cause the action in a non-determined way: according to the event-libertarian, given the past and the laws, it could have been the case that other reasons could have non-

deterministically (and nondeviantly) caused a different action (O'Connor & Franklin, 2019).

Non-causal accounts of free will

Non-causal libertarians contend that the power of self-determination need not be causally structured. Instead, an intentional action begins with a basic mental action, such as a decision or a choice. An action as e.g., raising one's arm is held to be a non-basic, complex action that is constituted by a basic action that brings about a certain bodily movement (Clarke & Capes, 2017). An influential defender of non-causal libertarianism is Carl Ginet (e.g., 1989, 1990, 2008). According to Ginet, we must understand the relation between intentions and actions when we are interested in explaining actions, and this relation has nothing to do with laws of nature (Ginet, 1989, p. 34) Instead, we control our basic actions, as e.g., choices, simply by having them, and by the fact that they are ours. Given that the event which is your choice is not determined *by anything else*, it is only you that ensure that what you do actually occur: "If an event is S's action the S (but, of course, no one else) can ensure its occurrence, determine *that* it occurs and thus *whether or not* it occurs, just by performing it" (Ginet, 1998, p.22).

Agent-causal accounts of free will

According to agent-causal views, the agent herself must play a causal role in self-determined actions. Some agent-causal libertarians integrate both the agent herself but also her reasons for action in what causes actions (e.g., Clarke, 2003) whereas others deny that an agent's reasons play a causal role in self-determined actions (e.g., O'Connor 2000). In O'Connor's view, a free decision must be caused by the agent, and it must not be the case that what the agent causes must be causally determined by prior events. Agent causation is another form of causation compared to mechanistic causation, according to O'Connor, which involves "the characteristic activity of purposive free agents" (O'Connor, 2000, p. 113) As with mechanistic causation, agent causation is, in O'Connor's view, also grounded in a property or a set of properties. This means that any agent having the relevant internal properties have it "directly within his power to cause any of

a range of states of intention delimited by internal and external circumstances" (O'Connor, 2000, p. 113). If we think of properties that ground mechanistic causation, there is a direct causal function in such circumstances where these properties produce certain effects. In contrast, properties that ground agent causation are not direct in that sense: instead, they are "choice-enabling" which means that they, in suitable circumstances, are properties that ground the agent's power to freely bring about, or freely cause, a certain intentional state (O'Connor, 2000, p. 113).

Clarke (2003) argues that agent-causation views as the one defended by O'Connor, that excludes reasons as causes, appear unable to account for why reasons explain actions. In order to avoid this problem, Clarke defends an "integrated" agent- causal view, according to which non-deterministic event causation is part of directly free actions. With such a combination, Clarke argues, we gain both the origination of action that is provided by traditional agent-causal accounts, but also what is needed for reasons-explanations (Clarke, 2003, pp. 135-136).

1.5.3 A disclaimer about how free will is to be discussed in this thesis

One of the premises in the Revision Argument is that according to folk psychology, we have a libertarian free will. However, what kind of libertarian free will that is required according to folk psychology is not specified, so in order to assess it, it could be argued that we must know more about what libertarian theory of free will that is intended. However, the objections towards the Revision Argument that are discussed in this thesis are not concerned with this particular issue. Instead, these objections reject the general claim that folk psychology requires libertarian free will and argue that a compatibilist view of free will is sufficient both for the folk psychological understanding of responsibility and desert as well as for the law. Hence, according to these objections, free will compatibilism can do the job that the Revision Argument claims has to be done by libertarian free will. This view, in turn, means that the ideas concerning free will and responsibility that will be discussed are compatibilist ideas, i.e., ideas

that accept the determinist hypothesis. Given the scope of this thesis, I will hence not go further into the specifics of each of the types of libertarianism.

1.6 Some theories of mental states

1.6.1 Introduction

In any theory of free will, a person's mental capacities play a fundamental role in explaining why and when a person exercises her free will. In order to make sense of this claim, we must have an idea of what mental states and mental capacities are.

In contemporary discussions of the nature of mental states, a central theme is to find a place for the mind in a world that is fundamentally physical (Kim, 2000, p. 2). As was noted in section 1.3, physicalism entails that mental states are, or are derivative from, physical states.

In the discussions to follow in this thesis, physicalism will be a constraint for what theories of mental states that are possible candidates in a theory of free will. In the following subsections, I will introduce some theories about mental states that are physicalist in the relevant sense.

1.6.2 The identity theory of mind

According to Jaegwon Kim (2000), the current debates on the mind-body problem can be traced back to Herbert Feigl's paper "The 'Mental' and the 'Physical' " (1958) and J.J.C Smart's "Sensations and Brain Processes" (1959).[9] Both these papers defend, independently of each other, an approach to the nature of the mind that has come to be known as the identity theory.

[9] Kim acknowledges that U.T. Place (1956) developed a similar idea some years before Smart and Feigl. According to Kim, Place's paper has not been as influential as Smart's and Feigl's. However, Smart's paper was a "refined and elaborated" version of Place's ideas (Chalmers, 2002, p. 4), and he is often mentioned as one of the earliest proponents of the identity theory (see e.g., Chalmers, 2002 and Smart, 2017).

The core idea of the identity theory is that mental states and processes are identical to states and processes of the brain. It does not entail that the brain is the mind in the sense that whatever one's brain weighs this mass is the weight of one's mind. Rather, the identity statement of the identity theory is to the effect that mental states *are* brain processes, as contrasted to being merely *correlated* with brain processes (Smart, 2017).

According to the identity theory, the meaning of two sentences must not be identical in order for the sentences to have the same reference. U.T Place discusses this relation as follows:

> Those who contend that the statement "consciousness is a brain process" is logically untenable base their claim, I suspect, on the mistaken assumption that if the meanings of two statements or expressions are quite unconnected, they cannot both provide an adequate characterization of the same object or state of affairs: if something is a state of consciousness, it cannot be a brain process, since there is nothing self-contradictory in supposing that someone feels a pain when there is nothing happening inside his skull. By the same token we might be led to conclude that a table cannot be an old packing case, since there is nothing self-contradictory in supposing that a table cannot be an old packing case, since there is nothing self-contradictory in supposing that someone has a table but is not in possession of an old packing case. (Place, 1956, pp. 45-46)

The trick is, according to Place, to keep distinct the "is" of composition and the "is" of definition. The claim "sensations are brain processes" makes use of "is" of composition, and hence it is an *empirical* matter whether this claim is true or false.

Smart and Feigl focus instead on the distinction between meaning and reference: "very bright planet seen in the morning" and "very bright planet seen in the evening" both refer to the same entity, Venus. It is a contingent matter if the references are identical. In a similar vein, "sensations are brain processes" is a contingent claim (Smart, 2017).

Paul Feyerabend (1963) argues that the identity thesis formulated as an identity relation between mental processes and brain states is problematic since it implies a version of dualism: "It not only implies, as it is intended to imply, that mental events have physical features; it also seems to imply (if read from the right to the left) that some

physical events […] have nonphysical features. It thereby replaces a dualism of events by a dualism of features" (Feyerabend, 1963, p. 295). Feyerabend's diagnosis is that this way of stating the identity is not sufficient for deciding the issue between monism and dualism. If we want to defend a monistic thesis, we must approach the matter differently: the identity theory should not only point to the fact that mental events are physical events, it also favors a redefinition of mental terms (Feyeraben, 1963, p. 296).

W.V.O Quine also endorses the idea that mental event terms appear superfluous in the light of the identity thesis, and that terms denoting the physical states that are identified with the mental states would be useful. In his view, the mental states do not exist anyway, but the physical states do: "[t]he bodily states exist anyway; why add others?" (Quine, 1960, p. 264).

One of the most powerful arguments against the identity theory is the multiple realizability argument, originally put forth by Hilary Putnam (1967). Putnam argues that in order for an identity theory to be true, a mental state, such as pain, must have some common physical/chemical basis in all kinds of creatures that can experience pain. But arguably, we want to be able to ascribe pain also to creatures that do not share our physical constitution. This criticism targets what is known as "type-type physicalism," according to which a certain type of mental state is identical with a certain type of brain state. A less demanding version of physicalism is "type-token" physicalism, according to which a specific type of mental events can be realized by different tokens of brain states. However, a type-token kind of physicalism is a much weaker thesis than type-type physicalism: it does not specify what kind of physical properties that are identical with mental kinds, and that seems to be problematic from an explanatory point of view: how can we form scientific laws that include mental states, if they are identical with a perhaps infinite set of physical properties?

David Lewis approaches the identity theory from this perspective, and changes focus from identity with a specific neural kind to the role of causality in the identification of mental states:

> The definitive characteristics of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states possess the definitive characteristics of experience, they must be the experiences. (Lewis, 1966, p. 17)

This view allows for the same kind of mental event to be instantiated in different physical systems, as long as it has the "most typical" causes and effects that are associated with this mental kind.

1.6.3 Eliminative materialism

Eliminative materialism, or eliminativism, is a radical position about the nature of mental states and the meaning of mental state terms, rooted in the views defended by Quine (1960) and Feyerabend (1963). According to eliminativism our common-sense view of the mind is fundamentally mistaken, and some or all of the mental states as they are understood by people in general, do not exist.[10]

Paul Churchland argues that an adequate and accurate theory of thinking and action should have considerable explanatory power, but folk psychology does not meet this requirement since it, for example, cannot explain why certain mental phenomena occur (P. M. Churchland, 1981, pp. 69-73). Another line of argument proceeds from the fact that folk theories change over time since they are commonly mistaken about all sorts of things. For example, the folk psychological understanding of certain mental disorders has undergone a radical ontological change in light of scientific discoveries about these disorders. Historically, some kinds of bizarre behavior were explained with reference to demons or other supernatural powers. In (most) contemporary accounts of behavior related to mental disorder, there are no references to supernatural powers: such explanations turned out to be empty since they referred to things that are not real (Ramsey, 2019).

[10] Ideas such as those defended by Quine and Feyeraben were preceded by C.D Broad in *The Mind and its Place in Nature* (Broad, 1925) and Wilfred Sellar's article "Empiricism and the Philosophy of Mind" (Sellars, 1956).

Eliminative materialism has been criticized from different perspectives. One is that the eliminative materialists ignore the explanatory power that folk psychology has when it comes to explaining and predicting behavior (see, e.g., Fodor, 1987). Other critics claim that folk psychology as an explanatory framework is committed to less than eliminativists typically assume. Instead, folk psychological explanations are not very ontologically demanding and are compatible with a wide range of ontological claims about the nature of the human mind (see e.g., Jackson and Pettit, 1990).

1.6.4 Property dualism

Property dualism holds that mental phenomena are non-physical properties of physical substances. Mental states are, hence, irreducible to physical states. One of the fundamental motivations for this approach is that reductive physicalism is an "unreasonably strong claim" (Hellman & Thompson, 1975, p. 551) More specifically, what is unreasonably strong is the claim that all scientific terms (in this case, mental terms that figure in scientific explanations) can be given explicit definitions in physical terms. Ernest LePore and Barry Loewer express their view on this subject as follows:

> It is practically received wisdom among philosophers of mind that psychological properties (including content properties) are not identical to neuropsychological or other physical properties. The relationship between psychological and neurophysiological properties is that the latter *realize* the former. Furthermore, a single psychological property might (in the sense of conceptual possibility) be realized by a large number, perhaps infinitely many, of different physical properties and even by non-physical properties. (Lepore & Loewer, 2011, p. 181)

Lepore & Loewer argue that mental states are multiply realizable in a "perhaps infinitely many" of different physical properties. As mentioned above, it was Hilary Putnam (1967) who first introduced the concept of multiple realizability in the mind-body discussion. As later writers have stressed, the multiple realizability of the mental is a conceptual point: it is an a priori conceptual fact about mental properties that their specification does not include constraints on the physical properties that realize them (Kim, 2002, p. 137). Since the

irreducibility of mental states is *conceptual* irreducibility, it concerns irreducibility of mental *types*, not mental *tokens*. In other words, it is impossible to reduce the mental type of e.g., pain, to certain physical properties, since pain can be realized by a perhaps infinitely many different physical properties. A response to the contention that mental properties are irreducible is to suggest that, e.g., pain is reducible to a *disjunction* of physical properties. This suggestion has been discarded by e.g., Jerry Fodor, who argues that the disjunction strategy makes mental states unable to figure in scientific laws. Since, Fodor argues, for a special science (like psychology) to be reducible to a physical theory, each kind in the special science must have a nomologically coextensive kind in the physical theory. In Fodor's vocabulary, *P* is a kind in a science just in case the science contains a law with *P* as its antecedent or consequent (Fodor, 2002, pp. 131-132). If mental properties are identified with a conjunction of physical properties, this conjunction of physical properties does not qualify as a kind, since a disjunction of heterogenous kinds is not a kind itself. And that would make mental states disqualified for being included in scientific theories, and hence, also disqualified from figuring in causal relations. If mental states lack causal power, it is, supposedly, connected to a number of problems in other domains. For example, it seems difficult to justify that we blame and punish people on the ground that certain mental states were involved in their bad actions. Kim argues that if we accept multiple realization of mental states, we must choose between either allowing disjunctive kinds, or acknowledge that our mental concepts do not pick out kinds and properties in the world (Kim, 2002, p. 147).

The discussion about whether mental properties can figure in causal explanations depends heavily on one's view of causation. Kim defends a "production view" of causation according to which an antecedent causes an effect if and only if it is both necessary and sufficient for the effect to come around, but many writers have objected that this view is problematic. As for example, Christian List & Peter Menzies (2017) discard the production view and argue that if we adopt a counterfactual approach to causation, it becomes obvious that mental causation cannot be reduced to physical causation. I will

return to questions of mental causation in chapter four and discuss List & Menzies' view to some extent in chapter five.

1.6.5 Functionalism

According to functionalism, what makes something a mental state of a particular kind– a belief, a smell, a pain – is the *functional relations* the state bears to the subject's other mental states and behavior (see, e.g., Block, 1994). If taking the mental state of pain as an example: even if the pain in you corresponds to a specific neural state *n1*, *n1* is not included in the definition of what it means that you are in pain. *n1 realizes* your pain but *is* not the pain. In another organism with a different neural system, the neural state *n2*a could realize a mental state with the same function. But since the mental state realized by *n2* occupies the same functional role in this system as pain does in you (let us say that pain makes you try to avoid the cause of the pain) we can conclude that this different organism can have the mental state of pain as well, even if this mental state is realized by radically different physical processes.

Functionalism is not a homogenous theory of mind. Different functionalist theories have been developed in relation to different aims. An early version of functionalism, developed as a response to the difficulties behaviorism faced, was machine state functionalism, perhaps most closely associated with Putnam (1975a, 1975c). According to this approach, the mind should be viewed as a probabilistic automaton. The theory aims to specify the probability with which the system, i.e., the mind, would enter a certain state, i.e., mental state, given a certain input, and then produce a certain output. Mental states are, in this framework, "machine table states" which are specified in terms of not only their relations to inputs and outputs but also to other states in the system at the time being (Levin, 2018).

Another functionalist approach is "psycho-functionalism" which is closely related to cognitive science. According to psycho-functionalism, associated with, e.g., Jerry Fodor and Ned Block (1972), psychology is an irreducibly complex science employing *purposive* explanations. Such explanations are also employed in biological sciences: as for example, in order to understand what a

heart is, we describe it as something that pumps blood, and as such it occupies a special role in the overall system of which it is a part. Analogously, mental states such as beliefs and desires are determined by the functional or causal role they have according to our best scientific psychological theory.

As mentioned above, Lewis (1966) defends a type-token kind of identity theory, in which mental states are identified by their causal roles, and the causal roles are most plausibly filled by neural states. Lewis' functionalism is called "analytic functionalism" and is specifically concerned with the meaning of mental state terms, and how they can be "translated" into functional descriptions that preserve the meanings of these terms. If this is a successful approach, Lewis' view escapes an objection to the original identity theory that charges it of implying property dualism, and hence, he succeeds in escaping the threat of epiphenomenalism facing the property dualists.[11] I will return to the Lewisian approach to causation in chapters four and five.

1.7 Summary

In this chapter the main themes underlying the argumentation of this thesis have been introduced and briefly discussed: neurolaw, free will and some different physicalist theories of mental states. Neurolaw was briefly described as a heterogenous, interdisciplinary research field that can be divided into the areas of assessment, intervention and revision. The Revision Argument, which is the point of departure for the discussions in this thesis, is a contribution to the theoretical part of the revision area of neurolaw. Some central features in free will compatibilism and free will incompatibilism were described. In the final section of this chapter I provided a brief description of a few theories of mental states that are relevant for the discussions in this thesis. In the next chapter, I will present and discuss the details of the Revision Argument.

---

[11] An objection that Smart (1959) attributed to Max Black (Block, 2007, p. 249) and is more recently defended by Stephen White (2007).

2 The Revision Argument

2.1 Introduction

The Revision Argument plays a central role in the discussion in this thesis. In this chapter this argument will be described and discussed in detail.

The Revision Argument is what I take to be the most plausible interpretation of an argument put forward by Joshua Greene and Jonathan Cohen in their seminal article "For the Law, Neuroscience Changes Nothing and Everything" published in 2004 and since then one of the most cited and discussed papers in the neurolaw debate.

The main claim in the Revision Argument is that from a philosophical point of view, we have reasons to think that we do not have free will, at least not in the sense that is presupposed in our everyday understanding of human action, as well as in legal contexts. The rapid development of neuroscience has provided further evidence supporting this view. Greene and Cohen describe the role of neuroscience in the free will debate as follows:

> [C]ontrary to legal and philosophical orthodoxy, determinism really does threaten free will and responsibility, as we intuitively understand them. It is just that most of us, including most philosophers and legal theorists, have yet to appreciate it [...] Neuroscience has a special role in this process for the following reason. As long as the mind remains a black box, there will always be a donkey on which to pin dualist or libertarian intuitions. For a long time, philosophical arguments have persuaded people that human action has purely mechanical causes, but not everyone cares for philosophical arguments. Arguments are nice, but physical demonstrations are far more compelling. What neuroscience does, and will continue to do at an accelerated pace, is elucidate the 'when', 'where', and 'how' of the mechanical processes that cause behavior. It is one thing to deny that human decision-making is purely mechanical when your opponent only offers a general, philosophical argument. It is quite another to hold your ground when your opponent can make detailed predictions about these mechanical processes work, complete with images of the brain structures involved and equations that describe their function. Thus, neuroscience holds

> the promise of turning the black box of the mind into a *transparent bottleneck*. (Greene & Cohen, 2004, pp. 1781-82)

The main concern in the Revision Argument is that certain parts of our legal responsibility practice are founded on libertarian intuitions about people's free will, while we have strong reasons to reject the belief in libertarian free will. Philosophical arguments against the view that people have a free will have been put forward many times throughout history, but now there are also physical demonstrations in support of the claim that human behavior is purely mechanical. We have, hence, both philosophical arguments and scientific evidence which urge us to revise the folk psychological understanding of human action. Since our current legal system(s) and the normative standards therein are grounded in a folk psychological understanding of behavior, they will be affected by such a revision, Greene and Cohen argue.

In order to see what kind of folk psychological explanations that the Revision Argument focuses on, consider the following example:

> Cornelia, who is out on a walk, approaches a child who is about to drown in a pond. Cornelia considers if she should help the child or not but chooses not to since that would require her to go out in the water and she doesn't like wet clothes. She continues her walk, and the child drowns.

Given that there is nothing wrong with Cornelia, i.e., she has no cognitive or volitional deficits due to a mental disorder that could be used in an explanation of her behavior, it seems uncontroversial to say that Cornelia acted wrongly and deserves to be blamed for what she did. It seems that Cornelia easily *could* have made the choice of saving the child and that not doing so was a selfish choice which led to horrific consequences. The assumption that Cornelia had a choice plays a vital role here, because it is plausible to assume that we would perceive the situation differently if Cornelia was somehow coerced to leave the child behind for some reason that we acknowledge as legitimate. For example, it could have been the case that Cornelia wanted to save the child, but someone held her back so she was physically unable to go out in the pond, and then she would not deserve blame for not saving the child.

The above line of reasoning is a typical folk psychological description of an action and its moral implications. The possibility to choose how to act hinges on different things. One such thing is that we are aware of the nature of our actions. For example, if I intentionally grab and keep someone else's wallet when she drops it without noticing it, we normally think I could have chosen not to do so. But if I by mistake take someone else's wallet because it looks exactly like my own, it seems like the choice to give the wallet back to the owner was not open to me, given what I believed in the situation at hand (that I was the owner).

In the legal context, the difference between intentional and unintentional acts plays an important role. When assessing whether a crime has taken place, it is not only the objective circumstances that are considered, i.e., the circumstances that can be objectively established regarding the behavior on the crime scene. The mental characteristics of the agent are also of vital importance in the assessment of whether a crime has been committed, and whether the defendant is to be viewed as *deserving* blame or punishment for what has happened.[12]

But if we assume that determinism is true, it seems that everything we do is equally determined to happen: regardless of whether we ourselves think that we can choose to act or not, the choice and the act are not "up to us." If we imagine a mechanical event, as, for example, a bowling ball hitting some bowling pins, there is nothing in that event that is "up to" the bowling ball. According to the determinist hypothesis, our actions are not more "up to us" than the whereabouts of a bowling ball (and the pins) are "up to the ball." We are of course more complex than bowling balls, but complexity does not play a distinguishing role according to this view: everything that happens do so out of necessity, and this fact holds for the simplest as well as the most complex systems in the universe.

Even though there are obvious differences between people and non-living things with regard to how they interact with the world, i.e.,

[12] In certain cases, people can be legally responsible and punished even if they didn't commit a criminal action intentionally. For example, a parent can be legally responsible on the behalf of her child, and a manager can be responsible for what her employees do. Such cases will be discussed in sections 2.4 and 2.5.

with regard to functional properties, one can ask if there is room for *desert* in a world where people's choices and action are no less determined than is a bowling ball's movements. According to the Revision Argument, the answer is no: there is no place for desert in such a world. Since retributive punishment requires desert in order to be justified, this view implies that retributive punishment lacks justification. And in so far as we demand *justified* legal punishment practices, we need to revise the legal system in the sense that we must remove the retributive principle from the system and replace it with a justification that does not rely on desert.

In this section I have provided a brief background of the moral and legal context in which the Revision Argument was developed, and I pointed out the main target of the argument. In the next section, 2.2, the argument is presented in the form of three premises and a conclusion, after which I will discuss the content of these premises in more detail in section 2.3. The Revision Argument holds that parts of our legal responsibility practice are unjustified. Responsibility, however, is not a clear-cut concept. In section 2.4, some different interpretations of this concept are presented. In section 2.5, I will make clear what notion of responsibility I use when I discuss the Revision Argument and the objections to it in this thesis. A summary of the chapter is provided in section 2.6.

2.2 Outline of the Revision Argument

The Revision Argument can be spelled out as follows:

> THE REVISION ARGUMENT
>
> (P1) **Punishment needs to be justified**
> One of the functions of the legal system is to punish those who engage in criminal behavior. Legal punishment should be based on a principle that makes the punishment morally justified.
>
> (P2) **The current retributivist justification**
> In the current legal system, the justification of punishment is at least partly retributivist. Legal punishment is legitimized by folk morality, and folk morality is based on a folk psychological understanding of behavior, according to which:

(a) people have, and their actions are explained partly in terms of, a libertarian free will.

According to folk morality, the following principles hold for a justified retributive punishment:

(b) it is only justified to punish those who deserve it, and

(c) someone deserves punishment only if she is responsible for her actions, and

(d) someone is responsible for her actions only if she can act freely, and

(e) someone can act freely only if she has a libertarian free will.

(P3) **Undermining the retributivist justification**
If we accept determinism, no one can have a libertarian free will, since actions are determined by earlier events and the laws of nature. The view that human action works according to deterministic and mechanical principles is supported by neuroscience since people's behavior can be accounted for in purely neuroscientific terms. In a neuroscientific explanation of action, free will is not required in order to make sense of why people act as they do. Hence, neuroscience supports the view that according to the best explanation, free will is not involved in human action. ('Neuroscience turns the black box of the mind into a transparent bottleneck'(Greene & Cohen, 2004, p. 1781).)

Conclusion
We have strong reasons to reject the hypothesis that people have a libertarian free will. However, libertarian free will is required by folk psychology and folk morality in order for people to be responsible in the sense that they can deserve punishment. Consequently, if there is no libertarian free will, we are never justified in punishing people for their actions on the ground that they deserve it. This means that accepting the principle according to which it is only justified to punish people if they deserve it, implies that it is never justified to punish people. (P2-P3)

Since retributivism is part of our current legal system, and it has turned out that retributivism cannot be a principle according to which we justifiably punish people, we need to revise the legal system in the sense that retributivism must be removed from the punishment practice, and replaced with another theory of punishment that is not based on the requirement that a person must have a libertarian free will on order to be justifiably punished. (P1)

2.3 Elaborating the argument

The central aim of this section is to specify how to understand the most significant claims and concepts in the Revision Argument, which is, in short, an argument to revise the law – deleting the retributive element – based on neuroscientific findings. At some points, this elaboration will be done by providing additional information from the original article by Greene & Cohen. At other points, concepts and claims are underspecified and/or open for more than one reading, and in these cases, I will specify how I understand them in the discussions to come. Premise two will be given the most attention due to its complex structure. It is also premise two that gains most criticism in the objections that are discussed in the ensuing chapters of the book.

2.3.1 First premise: Punishment needs to be justified

A basic feature of the legal system is that when people commit criminal actions, they can be legally punished for them. Punishment is not exercised on an ad-hoc basis, but under principles that guide when, why and how the punishment practice is implemented.

When I talk of "the legal system" in the discussions to come, I do not have a specific system in mind. Instead, it is best understood as a basic tenet of many legal systems; more specifically, I take my arguments to apply to any system that punishes at least partly with reference to retributive principles as justification, distinguishes between actions that are carried out freely and those that are not, and gains legitimacy from folk morality.

A starting point in the discussions to come, then, is that the kind of legal system that we are concerned with is legitimate to the extent that it is, to a sufficient degree, in line with, and justified by, folk moral intuitions. What people in general think is morally right and wrong matters to the legitimacy of the legal system. For example, if people, in general, feel that capital punishment is wrong, and capital punishment is an element in the legal system, then this element lacks

legitimacy.[13] However, premise one is concerned with moral justification, not legitimacy. There are different ways to understand what such justification consists in. In the discussions in this thesis, one of the questions is whether a certain theory of punishment (retributivism) provides a moral justification of punishment *that is in line with folk morality*. Or put in other words: is the retributivist justification of punishment consistent with folk moral intuitions? I interpret the claims of both the main proponents of the Revision Argument as well as the critics that I discuss in this book as follows: they agree that the moral justification of the legal system and punishment is to be sought in, or grounded in, folk morality. The moral justification question is intimately tied to legitimacy. The kind of legal systems that we are concerned with are legitimate to the extent that they are in line with folk morality.

Plausibly, the claim that punishment needs to be morally justified is embraced by folk morality. This means that according to folk morality, legal punishment must rely on a principle that morally justifies punishment.

An alternative way to conceive of the moral justification of the legal system would be to think of it as independent of folk morality. We could ask whether a certain moral justification of punishment is in line with *the correct* morality, rather than whether it is in line with the moral principles embraced by people in general. This question would presume that there are correct, or true, moral principles, but whether this is the case is not something that I will take a stand on here. Instead, as explained above, the assumption in the Revision Argument is that folk psychological and folk moral views are what make legal punishment legitimate, and the question is to what extent these folk moral and folk psychological assumptions are justified in

[13] The content of folk morality is complex, and to say that folk morality is what people in general think is morally right and wrong is a simplified way of characterizing it. There are different methods of assessing the content of folk morality and folk psychology, philosophers do often rely on their own intuitions in this regard, but there are also a number of experimental studies of these subjects. I will discuss such experimental studies in chapter three. In chapter six I will discuss a view of folk morality that allows for immediate responses to deviate from more well-reasoned moral beliefs, but where both the immediate response and the more well-reasoned belief can be called folk morality.

the light of other beliefs that people may be inclined to embrace, such as that the world is (sufficiently) determined.[14]

What makes it the case that a principle justifies, or fails to justify, punishment? As will be clear in the discussions to come, the moral justification not only depends on whether the principle in question is part of folk morality. It also depends on certain factual beliefs people have, and to what extent these factual beliefs are *epistemically* justified. Different moral principles refer to different states of affairs or properties of actions as the states of affairs and properties that make the actions right or wrong (i.e., morally justified or not). Consequently, whether a principle will succeed in distinguishing actions that are right from actions that are wrong, depends on facts about the world, e.g., whether the relevant properties are instantiated. For example, if we are to apply the moral principle that we should maximize the total sum of well-being, this principle presupposes that we can make interpersonal comparisons of levels of well-being. Investigating this is not a moral inquiry but a factual one, concerned with what the world is like. Likewise, if we want to apply the moral principle of retributivism, this principle presupposes that we can distinguish between those who deserve to be punished from those who do not, which, according to the second premise of the Revision Argument, depends on whether people have libertarian free will or not. This inquiry is not a moral one but a factual one, concerned with what the world is like.

To sum up: the first premise states that legal punishment needs to be morally justified, which here is to be interpreted as justified according to folk morality. This, in turn, means that legal punishment, to be justified, needs to conform to a moral principle or standard that is part of folk morality and that sets the criteria for why and when it is morally right to punish someone. Epistemic justification enters the picture in the assessment procedure of whether those criteria are met or not: we can be more or less epistemically justified in believing that

[14] This means that the methods used to explore what is part of folk morality, might be very similar to the methods of philosophers who describe their methods as ways to find the correct or most plausible moral principles. However, my interests are distinct from theirs.

the properties that according to folk morality are required for the justification of punishment are present or believing that they are not.

2.3.2 Second premise: The current retributivist justification

The second premise of the Revision Argument is concerned with the specification of how legal punishment is currently justified. The structure of this premise is complex, and in the following section I will briefly discuss the central concepts and clarify how they will be used.

Retributivism

(P2) states that the justification of legal punishment rests, at least partly, on retributivist grounds. Retributivism is traditionally viewed as a backward-looking, merit-based theory of punishment. As such, it seeks justification of punishment in a past action, rather than justifying punishment on the grounds of certain goods that will be gained in the future if the punishment is executed. Alec Walen describes the central features of retributivism as follows:

> [Retributive justice] is best understood as that form of justice committed to the following three principles: (1) that those who commit certain kinds of wrongful acts, paradigmatically serious crimes, morally deserve to suffer a proportionate punishment; (2) that it is intrinsically morally good – good without reference to any other goods that might arise – if some legitimate punisher gives them the punishment they deserve; and (3) that it is morally impermissible intentionally to punish the innocent or to inflict disproportionately largely punishment in wrongdoers. (Walen, 2016)

Retributivism captures a feature that might appear as intuitively morally plausible: that it is legitimate and intrinsically good to punish someone who has done wrong. Immanuel Kant provided an influential statement of the place of retribution in punishment in the 18th century:

> Punishment can never be administered merely as a means for promoting any other Good, either with regard to the Criminal himself or to Civil Society, but must in all cases be imposed only because the individual on whom it is inflicted *has committed a Crime*. For one man

> ought never to be dealt with merely as a means subservient to the purpose of another [...] (Kant, 1887, p. 195)

Duff & Hoskins (2018) label Kant's form of retributivism as *positive*. According to positive retributivism, an offender's wrongdoing provides *a reason in favor of* punishment. In contrast, according to *negative* retributivism, wrongdoing is rather a constraint for punishment, which means that punishment should be imposed only on those who have done wrong, and only in proportion to their wrongdoing – so, the punishment is thus *constrained* by the level of wrongdoing. One example of negative retributivism is, e.g., John Rawls' view: "What retributionists have rightly insisted upon is that no man can be punished unless he is guilty, that is, unless he has broken the law" (Rawls, 1955, p. 7).

Besides the difference between positive and negative accounts of retributivism, which concerns the issue of whether wrongdoing is a reason to punish or merely a constraint on who we may punish, there are several different ways to describe what a retributive theory is about, in a more specific way. According to John Cottingham, there are so many different definitions of "retributivism" that it is doubtful whether it is useful at all:

> Philosophers persist in talking of "retribution" and "retributive theory" as if these labels stood for something relatively simple and straightforward. The fact is that the term "retributive" as used in philosophy has become so imprecise and multivocal that it is doubtful whether it any longer serves a useful purpose. (Cottingham, 1979, p. 238)

A common way to describe retributivism as a "desert theory" according to which "punishment is meted out because it is deserved" (Golding, 1975, p. 89 cited in Cottingham p. 239). Cottingham's point is that even though retributivism basically is the thesis that it is morally permitted to punish those who deserve it, as contrasted to e.g., a theory according to which it is permitted to punish those who act wrongly because it has desirable consequences, the retributivist thesis can be further elaborated in different directions. For example, Cottingham mentions retributivism as a "repayment theory" according to which punishment is inflicted in order to make the

offender "pay" for his offense and re-establish *status quo* between members of the society (p. 238). Alternatively, retributive punishment is based on the idea that justice demands that an individual who has committed a crime should be punished according to the principle "an eye for an eye, a tooth for a tooth" (a version of this idea was the view defended by Kant) (p. 239).

Even if we might agree with Cottingham that the interpretation of the retributivist doctrine is far from uniform, it is at least clear that theories of retribution, regardless of what rationale is given as the reason for why punishment is justified, is a backward-looking kind of justification. Backward-looking theories of justification are traditionally contrasted to forward-looking theories of justification. The latter kind of theories are often called "consequentialist" or "instrumentalist" theories. Such consequences could, for example, be that punishment has a deterring effect on others, that we need to protect the members of society from criminals, or that punishment contributes to the stability of society.

Desert-based retributivism & basic desert

According to Berman (2011) the dominant classificatory framework of theories of punishment that had become orthodoxy by the latter decades of the 20th century is based on a distinction between consequentialist and retributivist justifications for punishment. In this framework, consequentialist theories justify punishment by the good that punishment produces, whereas retributive theories see punishment as justified by reference to the wrongdoers' supposed ill-desert: punishment is right if someone deserves it. Unfortunately, the definition of desert is not more straightforward than that of retributivism. Joel Feinberg (1970) analyzes desert as a triadic relationship between the agent who deserves something, that which is deserved and that which makes the agent deserve it. In the discussions to follow, the primary focus will be put on what it is that makes an agent deserve e.g., punishment. Feinberg calls this the "desert basis," which is the term I will use when I discuss the issue of on what basis people may deserve something. The person who deserves something is called the "desert subject" and that which is deserved the "desert object" (Walen, 2016).

If we adopt this terminology, a natural question that comes up concerns the desert object: what, specifically, could occupy this role? In other words: what is it that people deserve when they deserve something? As Berman puts it, "[t]ime and time again, it is said that, for the retributivists, 'punishment is justified because people deserve it' "(Berman, 2011, p. 437). It thus may seem as the desert object is punishment itself. But several scholars have argued that punishment should not be viewed as the desert object according to retributivism. Rather, the desert object must be something that can be achieved by punishment, such as, e.g., suffering (see e.g., Bagaric & Amarasekara (2000) and Duff (1990)). In response to such a view, Berman suggests that what wrongdoers deserve is something in line with that "their lives go less well" (Berman, 2011, p. 87). But, Berman continues, then it seems as if retributivism is no longer about punishment being intrinsically good, but rather that punishment is a way to obtain another value that we think of as an intrinsically good. And this thesis is, in turn, not substantially different from the consequentialist theory of punishment, according to which punishment is a way of attaining certain other values. Retributivism that is motivated in this manner makes the distinction between retributivist and consequentialist justifications of punishment less clear-cut. As Berman puts it, if retributivism is motivated thus, it has "morphed into an account that rests upon a justificatory structure that is plainly consequentialist" (p. 434). If this is correct, retributivism motivated by referring to certain kinds of desert-objects should be viewed as a subtype of consequentialist justifications for punishment: a "retributivist consequentialism" that can be meaningfully contrasted with varieties of "non- retributivist consequentialism."

The question of whether retributivism is, or could be, understood as a form of consequentialism will be discussed to some extent in chapter 7. However, for the main discussion in this thesis it will not be of any essential difference if we regard retributivism as a form of consequentialism in the sense Berman suggests, or as a deontological

theory of punishment, as long as desert is required for the justification of retributive punishment.[15]

Michael Moore defines of the core meaning of retributivism like this: "what is distinctively retributivist is the view that the guilty receiving their just desert is an intrinsic good" (Moore, 1993, p. 15). As Moore puts it, it is "intrinsically good" that the guilty receive their just desert. If we do not want to define retributivism along the lines of a certain value, we can put it in deontological terms instead: it is morally right to give the guilty person her just desert. Regardless of whether we consider the thesis of retributivism to be a deontological or consequentialist thesis á la Berman, retributive punishment cannot be justified if people do not deserve anything in a relevant sense.

When desert is discussed in the following chapters, it will be related to actions: someone can deserve something in virtue of performing an action with a certain moral status. This view of desert corresponds to what Derk Pereboom's calls "basic desert":

> The desert at issue here is basic in the sense that the agent would deserve to be blamed and praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely in virtue of consequentialist or contractualist considerations. (Pereboom, 2014, p. 2)

In a similar spirit, Saul Smilansky writes:

> My understanding of desert will in general reflect common usage. In its broadest sense, desert may imply a general idea of justice, roughly equivalent to that which is "due" to people. It is instructive for the importance of desert to note that desert in the broad sense can stand for all of justice in common speech. Here, however, we will be concerned with desert in a narrower sense. In this narrow sense, to say that A deserves X is to say that A is in certain respect due X (treatment or situation) on account of "intrinsic" features of himself or his acts. Desert in the narrow sense is intimately connected with the person. This is admittedly vague, but such vagueness seems to be part of the

[15] According to deontological theories, an action (as for example the action of punishing someone) is morally assessed not in relation to what desirable effects they bring about, but with regard to if the action conforms to a certain moral norm (see e.g., Alexander & Moore, 2016).

> concept. Desert is typically backward-looking, it is concerned with what a person deserves to get for what she has done. When we enquire about desert we are not asking, like consequentialists, what it would be more useful to everyone, in the future, to give her, but focusing only on what she should get herself [...] given a narrow concept of desert, we can see that a person's desert might differ from his entitlement, rights, or from what he ought to get from a consequentialist perspective. (Smilansky, 2000, pp. 13-14)

On the relevant concept of desert, then, to say that the agent deserves to be treated in a certain way (or deserves "to get X") for having performed an action, is to say that she is due that treatment (or to get X) just because she has performed that action. It is "basic" desert in the sense that it depends on no further moral justification of treating her in the relevant way, such as a consequentialist justification in terms of that treating her (or people like her) in that way has good consequences, or that there are conventions or contractualist considerations that make her entitled to be treated in that way. Rather, she deserves the treatment just because she has performed the action. Basic desert seems, hence, to be the notion of desert that is relevant for retributivism since retributivism is the idea that punishment is to be justified on backward-looking considerations alone. I will assume that the kind of desert referred to in the Revision Argument as well as the objections to it, is basic desert. When I discuss desert in the following chapters, I will have basic desert in mind, if I do not explicitly indicate otherwise.[16]

Smilansky further writes that a person's desert in the relevant sense depends on "[…] 'intrinsic' features of himself or his acts. Desert in the narrow sense is intimately connected with the person". As I interpret this claim, it concerns what it takes to be someone who can deserve punishment in the sense of basic desert. Some people – for

[16] In contrast, *non-basic* desert is ascribed to a person not only based on her actions (and their moral status), but on the basis of, e.g., consequentialist (or contractualist) considerations. Hence, non-basic desert attribution may be justified in cases where basic desert attribution is not. Such a view is, for example, defended by Henry Sidgwick, who suggests that desert claims can be justified with appeal to considerations about the values of consequences (Sidgwick 1907, p. 284). For more elaborated views on desert, see e.g., Feinberg (1970) and McLeod(2013). For an interesting exploration of different forms of desert, see Cupit (1996).

example small children and severely mentally disordered persons – plausibly do not deserve to be, for example, punished in the basic desert sense even if they have performed bad or illegal actions. Whether someone is a person who can deserve something in the basic desert sense depends, hence, on intrinsic properties of the person such as her mental states and capacities. This is where the relation between desert on the one hand and responsibility and/or free will on the other hand enters the scene.

Responsibility and free will

As formulated in (P2c), a person must be responsible for her actions in order to deserve punishment for what she has done. As mentioned above there are different notions of responsibility. In the Revision Argument, the two most obvious notions of responsibility are moral and legal responsibility. Even though legal responsibility in many aspects is different from moral responsibility, there are intimate connections between them. In the same way as the justification of legal punishment must correspond to the folk moral justification of punishment in order to be legitimate, legal responsibility practices are also legitimized by folk moral intuitions. If people in general think that the legal responsibility practice is morally unjustified, this legal responsibility practice lacks legitimacy among the people.[17]

(P2d) and (P2e) are both concerned with free will. To be responsible, it is required that one acts freely (P2d), and one can only act freely (in the relevant sense) if one has a libertarian free will (P2e).

---

[17] People's moral intuitions are of course not the only thing that determines how legal practices are designed. The point I wish to make is that the general justificatory basis upon which people are held legally responsible must, in order to gain legitimacy, be supported by folk moral intuitions. This general justificatory basis may, for example, be that someone must have certain capacities in order for being responsible in the relevant sense. Details of what people can be held responsible *for* is another issue, and it is plausible to think that such details may be more or less supported by folk moral intuitions, and that lack of folk moral support in some cases are not of direct relevance for the legitimacy of that particular aspect of the legal system. Another way to put it is that it seems reasonable to think that folk morality allows for some unsynchronized aspects of the legal system and folk morality, as long as the fundamental moral values as well as the overall system is generally supported.

According to (P2a), it is a folk psychological assumption that people have a libertarian free will, and that this can be part of the explanation of their actions.

As was shortly addressed in chapter one, there are different accounts of libertarian free will, and (P2) gives us no clear view of how the folk psychological account of libertarian free will should be interpreted. The point made in (P2), however, is that libertarian free will is incompatible with a determinist world view. When I mention libertarian free will in the discussions of this thesis, I am referring to an account of free will that holds that free will is not compatible with a fully deterministic universe.[18]

As previously mentioned, retributive punishment means that punishment is only permitted if it is deserved. I will in the discussions to follow talk about desert as a moral property a person can have in virtue of her intrinsic properties and her actions. Admittedly, it may sound odd to talk about desert as a moral property, but as I think of it, this is a way of expressing that someone is, in virtue of her intrinsic properties, a person that has the moral property of deserving punishment just because she has performed an action with a certain moral status. Furthermore, desert is attributed based on backward-looking, moral considerations, and as such, fundamentally different from consequentialist considerations.

According to P2c, someone deserves punishment only if she is responsible for her actions. The notion of responsibility will be further elaborated upon in section 2.5, in order to specify what "responsibility" means in the discussions to come.

[18] I do not claim that libertarian free will is an untenable view, or that it is incompatible with a scientific explanation of human action. I remain neutral regarding the plausibility of libertarian free will in this book. The reason for why libertarian free will is left out of the discussion is of pragmatic reasons: in this book, I focus on discussions in which determinism is accepted. I discussed the content of the deterministic thesis and certain problems with this view in section 1.4.

2.3.3 Third premise: Undermining the retributivist justification

According to premise two, libertarian free will is necessary for retributive punishment to be justified. But the very possibility of free will in the libertarian sense is contested in philosophy: it is argued that such free will is incompatible with our world. In premise three, we are faced with the claim that neuroscience provides further evidence to this position.

According to Greene and Cohen, what neuroscience adds to the debate about the plausibility of libertarian free will is not any new, revolutionary facts about the mechanisms of the human brain and how it causes behavior. Rather, neuroscience provides more detailed descriptions of the "when," "where," and "how" of the mechanical processes that causes behavior, which makes it harder to hold on to the belief in libertarian free will for those who have not yet left it behind on the grounds of purely philosophical arguments:

> It is one thing to deny that human decision-making is purely mechanical when your opponent offers only a general, philosophical argument. It is quite another to hold your ground when your opponent can make detailed predictions about how these mechanical processes work, complete with images of the brain structures involved and equations that describe their function. (Greene & Cohen, 2004, p. 1781)

In light of this quote, (P3) is best understood as holding that neuroscience contributes further evidence supporting the hypothesis that there is no libertarian free will. The evidence Greene & Cohen have in mind is that the increasingly detailed neuroscientific explanatory framework of human behavior will make it possible to give precise explanations and predictions of behavior and how it is connected to the brain. And since neuroscientific explanations describe behavior and its relation to the brain without including any references to libertarian free will, libertarian free will is *not needed* in order to understand why people act as they do. Although there are several established philosophical arguments for why there is no libertarian free will, philosophical arguments do not always bite. Greene & Cohen think that even though there are people who resist philosophical arguments to the effect that there is no libertarian free

will, it is more difficult to reject detailed scientific evidence that supports this view. All in all, neuroscience provides support to the philosophical argument against libertarian free will. [19]

The question of how neuroscientific evidence can and cannot be used in order to find out things about free will is widely discussed, as is the question of to what degree neuroscience can contribute to our understanding of why people act as they do. These issues are addressed in the remaining chapters of this book. But before that I will discuss the conclusion of the Revision Argument.

2.3.4 The conclusion

The Revision Argument claims that we are not justified in punishing people on retributivist grounds. If retributivism plays a central role in our punishment practice and restricts it in the sense that we are never justified in punishing offenders *if they do not deserve it,* this creates a problem with regard to premise 1: that legal punishment should be based on a principle that makes the punishment morally justified. If we accept premise one, our legal system must be revised in the sense that retributivism must be replaced with a principle that can justify punishment also in the absence of libertarian free will.

This conclusion is contested. Most people agree that a legal system is needed for a functioning society, but if we accept the Revision Argument, we cannot practice one that legitimizes punishment according to retributivist principles. The main aim of this thesis is to analyze some serious objections to this conclusion.

Before that, however, it is worth saying a bit more about responsibility. The claim that we cannot build our legal system upon a retributive justification of punishment might steer up worries about responsibility: is the conclusion of the Revision Argument also that no one is ever *really* responsible? Responsibility is a central notion in the Revision Argument, but its meaning is not clearly defined. In the

[19] Another perspective to the comparison of folk psychological and neuroscientific explanations of behavior is that neuroscientific explanations of behavior is superior to folk psychological explanations in virtue of Ockham's razor: if we have two explanatory frameworks that explains the same phenomena and are equally successful, we ought to choose the framework that needs to postulate fewer ontological entities.

following section, I will discuss some different notions of responsibility that illustrates how the moral significance of responsibility can vary depending on which notion of "responsibility" one has in mind.

2.4 Different notions of responsibility

A worry that might arise in relation to the conclusion of the Revision Argument is that no one is ever responsible for his or her actions. As will become clear when we distinguish between different notions of responsibility, although responsibility is often central in discussions of free will and desert, which are in focus in this thesis, there are notions of responsibility that remain unaffected by such discussions. Therefore, to reject free will and desert must not imply a rejection of us ever being responsible.

There are many suggestions on how to distinguish between different notions of responsibility. Recent work has been done by e.g., Nicole Vincent (2011), and Ibo van de Poel (2011). Many of the contemporary taxonomies of notions of responsibility are influenced by the division made by H.L.A Hart, who in the following quote illustrates how we can talk about responsibility can be used in a number of different ways:

> As captain of the ship, X was responsible for the safety of his passengers and crew. But on his last voyage he got drunk every night and was responsible for the loss of the ship with all aboard. It was rumored that he was insane, but the doctors considered that he was responsible for his actions. Throughout the voyage he behaved quite irresponsibly, and various incidents in his career showed that he was not a responsible person. He always maintained that the exceptional winter storms were responsible for the loss of the ship, but in the legal proceedings brought against him he was found criminally responsible for his negligent conduct, and in separate civil proceedings he was held legally responsible for the loss of life and property. He is still alive and he is morally responsible for the deaths of many women and children. (Hart, 1968, p. 212)

Hart distinguishes between four different notions of responsibility: role-responsibility, causal-responsibility, liability-responsibility and capacity-responsibility. Below I will provide a short presentation of

each of these notions. Then, I will go on to discuss some connections, overlaps and contradictions that can arise between them. This list is not meant to be an exhaustive examination of all kinds of responsibility notions we can think of, nor is it the only taxonomy of different notions of responsibility in the literature. However, Hart's distinction will be sufficient for the purposes of this section, which is to illustrate some different notions of responsibility and discuss to what extent they are morally significant, and also to determine which notion of responsibility that is in focus in the Revision Argument.

2.4.1 Role responsibility

The captain is responsible for his ship, parents are responsible for the upbringing of their children, a doctor is responsible for the treatment of her patient, a bus driver is responsible for driving the bus to the right destination. All these examples connect a person X to a certain duty Y. This kind of responsibility seems to be closely connected to social organization and normative to its nature in one sense: according to a certain normative structure, X is expected to act in such a way as to fulfil the goals associated with Y. Hart thinks, even though he confesses not being sure about it, that what distinguishes duties viewed as included in the role-responsibility of a person from short-lived duties of a simple kind is that role-responsibility duties require care and attention over a protracted period (Hart, 1968, p. 213). However, in the present context, we need not care about the sharp distinction between duties that persist over time and more short-lived duties. Role-responsibility duties seem to be part of a more or less abstract bundle of duties with varying significance in relation to fulfilling the overall responsibility task. To be a "responsible person" in a role-responsible sense is something like being a person who cares about her duties, and make an effort to fulfill them.

2.4.2 Causal responsibility

In some contexts, "X is responsible for Y" means that X is (or, a significant part of) the cause of Y, as in "the fire was responsible for the animals panicking," or "the stormy weather was responsible for many car crashes." When "responsible" is used in this way, in relation

to someone's action, it is used to point out that a person's action was particularly significant in the production of the outcome in question; but there is no particular moral judgment connected to the action. Hart notes that there is a difference between the claims "*is* responsible for" and "*was* responsible for" about a past action: to say that someone *was* responsible for a car-crash indicates that it is causal responsibility one has in mind, as compared to saying that someone *is* responsible for a car crash, in which it is not merely that causal relation to the car-crash is indicated. However, to say that someone no longer living was responsible for a car-crash can indicate both causal responsibility and liability (Hart 1968, p. 215).

2.4.3 Liability

As already mentioned, "liability" is perhaps a more familiar term than "responsibility" in the legal domain. Even if we might view liability as a form of responsibility, Hart points out that there is a notion of legal responsibility that should not be viewed as synonymous with legal liability (1968, p. 217). According to Hart, to say that someone is liable is not necessarily to pick out anything else than the fact that someone "has to pay" in some sense (by paying in an ordinary sense, or be legally punished in some other way) for a damage that has been done and that can be traced to that person in some way – not necessarily morally. A person may be liable to pay compensation for harm caused by others, e.g., by her employees, even if she was unaware of her employees' whereabouts. To say that e.g., a company is legally liable for its products means (again roughly) that the company has to compensate the customer if it turns out that the product was defect or dangerous in some way. All these uses of "liability" focus on what has to be done in order to "restore the balance" without taking into consideration someone's actual blameworthiness.

In order to determine whether someone is legally liable or not for a certain action, one of the most basic questions is whether the action is considered criminal according to the law. If it is not, it is not possible to be legally liable for it, even though the action might appear as morally wrong. If the action is considered unlawful, a person is liable only insofar as she fulfills certain criteria at the time of action.

The assessment of whether someone does or does not fulfill these criteria is concerned with a more specialized range of topics, mainly with psychological conditions in the agent.

Hart asks: "How far can the account given above of legal liability-responsibility be applied *mutatis mutandis* to moral responsibility?" (Hart 1968, p. 225). As he thinks of it, in the moral domain "deserving blame" or "being blameworthy" will have to be substituted for "liable to punishment," and "morally bound to make amends or pay compensation" be substituted for "liable to be made to pay compensation." The assessment of whether someone is blameworthy or morally bound to pay compensation is made in basically the same manner as in cases with legal liability: it has to do with "a man's control over his own conduct, or to the causal or other connection between his action and harmful occurrences, or to his relationship with the person who actually did the harm" (Hart, 1968, p. 226). The important difference between legal and moral responsibility is not due to the meaning of the terms themselves but is rather due to different content of moral and legal rules and principles. In both legal and moral cases, the criteria of responsibility are restricted to "the psychological elements involved in the control of conduct, to causal or other connexions between acts and harm, and to the relationships with the actual doer of misdeeds" (Hart, 1968, p. 227). But there is a difference in what kinds of specific criteria that fall under these general ones. A legal system can hold people responsible in ways we condemn as deeply unjust, but, as Hart notes, "there are no conceptual barriers to be overcome in speaking of such a system as a legal system" (p. 227). In the moral case, there are such conceptual barriers:

> The hypothesis that we might hold individuals morally blameworthy for doing things which they could not have avoided doing, or for things done by others over whom they had no control, conflicts with too many of the central features of the idea of morality to be treated merely as speculation about a rare or inferior kind of moral system. It may be an exaggeration to say that there could not logically be such a morality or that blame administered according to principles of strict or vicarious responsibility, even in a minority of cases, could not logically be moral blame; none the less, admission of such a system as amorality would require a profound modification of our present concept of morality,

> and there is no similar requirement in the case of law. (Hart, 1968, p. 226)

Hart's description of the difference between legal and moral responsibility is instructive. Legal and moral responsibility have different criteria, and are attributed of different reasons. But equally important as it is to recognize them as separate from each other, is to recognize the relation between them. This point will be discussed in chapter four and more elaborated on in chapter six.

2.4.4 Capacity responsibility

We do not only talk about responsibility for specific actions, but sometimes we also distinguish between people who are responsible for their actions in a general sense, from people who are not. Such general claims – that someone is a person who is responsible for her actions – often refer to the fact that the person has certain psychological capacities that are viewed as necessary for being included in the group of people who are responsible for their actions, morally or legally (Hart 1968, p. 228). Children under a certain age are usually thought of as not being responsible for what they do, neither morally nor legally, and people with severe psychological disabilities are in most legal systems viewed as having at least an impaired capacity to take responsibility for their actions.

One reason to have a legal system that takes such capacities into account is that the efficacy of a legal system depends on that a sufficient number of those whom it concerns understand what the law requires of them (Hart 1968, p. 230). A legal system with strict responsibility, i.e., in which everyone could be held responsible for everything, including small children and people with mental disabilities, would be very demanding for those trying to secure that the rules within the system are followed, and would be costly due to the number of people that would be targets of punishment. It is probably a fairly uncontroversial idea that a legal system is (at least partly) a way of regulating human conduct through communication. If people cannot understand the laws, and therefore cannot adapt their behavior in order to avoid being punished, this criterion of communication has failed.

Another reason why a legal system restricts liability to people with certain capacities is the idea that one should only punish those who are actually morally responsible – in a sense that entails blameworthiness – for their actions. Having a legal system that restricts liability to people with certain capacities, is thus a way of trying to ensure that only those who have the capacities necessary for being morally responsible for their actions (in this sense) are punished by the legal system.

It is important, however, for the discussion to come in this thesis, that we keep two things apart. As was pointed out in the previous section, we can distinguish between legal liability on the one hand, and the corresponding moral notion – i.e., moral responsibility that entails blameworthiness – on the other hand. In the same way, we should keep distinguished two forms of capacity responsibility apart: the capacities required for legal liability, on the one hand, and the capacities required for moral blameworthiness, on the other hand. A central assumption in this thesis is that the capacities required for legal responsibility are at least roughly corresponding to the capacities that are required for moral responsibility. Hence, when we discuss the criteria for legal capacity responsibility (i.e., what capacities and properties are required for legal responsibility) these criteria will to a large extent coincide with the criteria for moral capacity responsibility.

2.5 The notion of responsibility in the Revision Argument

In the foregoing section I presented four different notions of responsibility: causal responsibility, liability, role responsibility and capacity responsibility. In this section I will discuss how these notions relate to the kind of responsibility that is the target of the Revision Argument.

In the Revision Argument, responsibility is a crucial part of premise two (P2). According to P2, a person must be responsible in order to deserve punishment, and she is responsible only if she acts freely, which in turn requires libertarian free will. Hence, according to the Revision Argument, libertarian free will is a necessary requirement

for being responsible in the sense that someone must be in order to deserve e.g., punishment.

As causal responsibility was characterized in the previous section, it does not require free will of any sort and must not be connected to desert. Someone may cause an event as a complete accident. Moreover, it sounds perfectly sensible to say that someone was causally responsible for an accident but is neither morally nor legally responsible for it (as for example, someone can be causally responsible for an accident, but it was not her fault, but due to a construction failure of the car.) To be liable, in the sense discussed above, is to be a person who, according to the law, should pay, or otherwise compensate, for something that has happened. But liability need not be connected to a freedom-condition (i.e., that the agent acted freely): people can be legally liable without being personally involved in the action or state of affairs for which they are held liable. This can, in turn, be explained with reference to their role responsibility: a parent may be legally liable for the whereabouts of her child in virtue of her being the parent of the child. Likewise, a manager can be legally liable for the wrongdoings of her employees in virtue of her role as a manager. The question of whether the parent, or the manager, have libertarian free will is not directly relevant to whether they have the relevant role-responsibilities for their children or employees. If the parent, or the manager, deserve punishment, they do not deserve it because of their own actions but because of the role they occupy.

Hence, role-responsibility, causal responsibility, and legal liability are not the relevant sorts of responsibility in the Revision Argument: we can use these notions of responsibility without assuming anything about free will. I will argue that the notion that is implied in the Revision Argument is that of *capacity* responsibility.

An illustration of why it is reasonable to understand responsibility as it is used in the Revision Argument as capacity responsibility is that, as noted above, in most legislations, a necessary condition for being held criminally responsible for one's actions is that one has *the right psychological capacities*. For example, children under a certain age are usually thought of as not being responsible for what they do (neither morally nor legally) and people with severe psychological disabilities

are in most legal systems viewed as having at least an impaired capacity to take responsibility for their actions. In the Revision Argument, there is no explicit mentioning of the requirement of certain capacities in order to be responsible. However, premise two states that responsibility requires libertarian free will, and libertarian free will can be understood as a capacity that one can have, or lack, in the sense that one can have the capacity to act out of one's free will, or lack that capacity.[20]

As will be clear in the discussions to follow, the claim that the capacity to act out of one's libertarian free will is a necessary requirement for legal retributive punishment is contested. In most legislations, certain cognitive capacities are specified as necessary for being liable, but it is doubtful whether any legislation specifies libertarian free will as a criterion for legal responsibility. Instead, it may be argued that other cognitive capacities are required for someone to be "capacity responsible", as for example, the capacity of being reasons-responsive.

Whatever one think is the required capacity for being capacity responsible, a further question is: when someone is capacity responsible and performs an action for which she is held responsible, what does this mean? Since the Revision Argument is about retributive punishment, this discussion concerns *what capacity that is required in order for retributive punishment to be justified.* As was discussed in section 2.3.2, retributive punishment requires desert, and more specifically, what Pereboom (2014) calls "basic desert". As previously noted, I will talk about basic desert as a moral property a person can have. Consequently, what is in focus in this thesis is what capacity a person must have in order to be someone who can acquire the moral property of basic desert. This moral property is, in turn, required in order for retributive punishment to be justified.

Hence, when we say that retributive punishment requires responsibility, this means that if someone is to be punished on retributive grounds, she must have the mental capacity that is required

[20] As was noted in section 1.5.2, libertarian free will is not a homogenous theory of free will but can be characterized in many different ways. Common for all of them is that they are not compatible with determinism.

for acquiring basic desert if she performs an action with a certain moral status. If someone performs such an action, and has the required capacity, she is morally responsible for the action in the sense required for being the proper target of retributive punishment. In other words, retributive punishment requires, in order to be justified, that the person who is punished is responsible in the basic desert sense.

Gregg Caruso & Stephen Morris label this specific kind of responsibility "retributive desert moral responsibility". (Caruso & Morris, 2017, pp. 840-841). When I discuss moral responsibility in this thesis, it is this notion that I have in mind, if I do not explicitly state something different.

Having made this point, I do not dismiss claims to the effect that there are other notions of moral responsibility: to me, it seems plausible that there are notions of moral responsibility that are not essentially connected to desert, and that these notions can serve as justification for *non-retributivist* punishment. The view that moral responsibility and justified (non-retributive) punishment can be had without accepting that people deserve anything in the sense of basic desert has been embraced by several philosophers, as for example J.J.C Smart (1961), Daniel Dennett (1984b), Saul Smilansky (2000) and Derk Pereboom (2014). David Hume can also be interpreted as defending this view (Russell, 1990, p. 560). In such a responsibility practice, what makes moral responsibility attribution justified, as well as the subsequent response (blame, praise, punishment, reward, etc.) is that this practice has consequences we consider as valuable. This kind of moral responsibility, its relation to folk morality and its applicability in the legal system, will be briefly discussed in chapter 7.

2.6 Summary

The Revision Argument holds that we should discard retributivism as a principle of punishment according to which we punish people in the legal system. The central feature of retributivism, as it is understood in this analysis, is that it is morally permitted to punish someone only if she deserves it.

When desert is discussed in this thesis, it is basic desert I have in mind. I will talk about basic desert as a moral property someone can have, and this means that she has the property of being such a person that deserves punishment in the basic desert sense. Basic desert can be contrasted to a notion of desert according to which someone has the property of deserving punishment because of consequentialist or contractualist considerations.

As I understand the Revision Argument, it is basic desert that is required for the kind of responsibility necessary for retributivism. Following Caruso & Morris, the kind of moral responsibility that I take to be the relevant in the Revision Argument is what they call "retributive desert moral responsibility." This notion is also the one that I will have in mind when discussing responsibility in the following discussions, if I do not state otherwise.

The crucial claim of the Revision Argument, then, is that we are not morally responsible in the sense of retributive desert moral responsibility, because in order to be morally responsible in that sense it is required that we have a libertarian free will – which we, most probably, do not have. If this is correct, the justification of legal retributive punishment is lacking.

This line of reasoning is much contested for various reasons. A common objection is that the conclusion that no one can deserve anything if she lacks libertarian free will is a misconstruction about what kind of free will is required for desert. With regard to legal retributive punishment, free will compatibilism provides a notion of free will that justifies retributive punishment. This argument is the topic of the next chapter.

3 First objection: Legal retributive punishment does not require free will

3.1 Introduction

According to Stephen Morse, criminal law "is a thoroughly folk psychological enterprise that is completely consistent with the truth of determinism or universal causation […] to be sure, criminal law doctrine and practice are also fully consistent with metaphysical libertarianism" (Morse, 2013a, pp. 27-28). In several different publications (e.g., 2004, 2006, 2007, 2013a, 2013b) Morse argues for the view that neuroscience has much more limited relevance for the law than "neuro-enthusiasts" think. "Neuro-enthusiasts" hold, according to Morse, that neuroscience shows that what the law presupposes about human behavior is flawed: they claim that the law is built on assumptions of the existence of a libertarian free will and a genuine choice, and that neuroscience strongly suggests that there is no such freedom since brains work in accordance with the same physical processes as any other mechanism in the world, which most often is assumed to be either deterministic or probabilistic. Neither of them leaving any space for libertarian free will. The neuro-enthusiasts do, hence, reason along the lines of the Revision Argument. In Morse's view, this criticism is due to a misunderstanding of what the legal system is about. [21] What the law requires in order to hold people responsible is nothing like a libertarian free will. Rather, Morse claims, what the law is concerned

[21] As was explained in section 2.3.1, when I talk of "the legal system" in this thesis, I do not have a specific system in mind. Many people in this discussion, for example Morse and Pardo & Patterson, are primarily discussing the U.S. legal system. When I talk about "the legal system", I think of a legal system that punishes at least partly with reference to retributive principles, that distinguishes between actions that are carried out freely and actions that are not, and also gains legitimacy from folk morality and folk psychology, as they are discussed in this thesis.

with when it comes to who deserves punishment is a conception of "rational personhood." The law assumes that people have mental states like intentions, beliefs, desires, and plans, and that they act in certain ways given these mental states. This view presumes that "people are *practical reasoners*, the sort of creatures that can act for and respond to reason" (Morse 2013, p. 31, my italics). Given this picture, the law is an action-guiding enterprise which provides rules and standards that guide people in their reasoning of how they should behave (Morse, 2013, p. 31). Morse continues: "Virtually all criminals are rational, responsible agents, and in such cases, punishment is premised on the basis of desert. No agent should be punished without desert for wrongdoing […] we cannot detain them unless they deserve it, and desert requires culpable wrongdoing" (Morse 2013, p. 29).

Nothing in this picture requires the existence of a libertarian free will or is incompatible with determinism, according to Morse. To believe in metaphysical libertarianism is, in Morse's view, "extremely implausible in the modern, scientific age:"

> Human beings, as complex as they are, are still part of the physical universe and subject to the same laws that govern all phenomena. In short, I believe that libertarianism does not furnish a justifiable foundation for an institution that is essentially about blaming and punishing culpable agents. If determinism or something quite like it is true, as I assume it is, then only compatibilism provides a secure basis for criminal responsibility. (Morse 2013, p. 28)

Law is, according to Morse, based on the folk psychological notion of rational personhood. This claim entails that the law's underlying understanding of the human mind is, in its essence, a folk psychological theory. Furthermore, according to Morse, but in contrast to what is maintained in the Revisionist Argument, the folk psychological view of free will is *not* libertarian:

> [O]n the most natural reading of what ordinary people mean by claiming that free will is foundational, [libertarianism] is not necessary even for them, and determinism is not inconsistent with current responsibility doctrines and practices […] I believe that most confuse libertarian free will with freedom of action, the ability to do what one wants. I recognize that one interpretation of hard determinism is that agents cannot do what they want because there is no alternative

> possibility other than the action that was performed. But this position is contested, and it is entirely consistent with determinism that people can act freely in the sense of doing what they choose to do based on their reasons for action and doing so without compulsion. (Morse 2013, p. 39)

Morse holds that libertarianism is an untenable view of free will from a philosophical perspective, and he also maintains that "ordinary people" do not think that a libertarian free will is needed in order to justify current (everyday and legal) responsibility practices and doctrines. What people do think is required for attributing responsibility to others is that they can act freely and also have the ability to do what they want. But these requirements, Morse argues, can be satisfied without a libertarian free will. As was already mentioned in section 1.5, there are different forms of libertarianism and some of them are compatible with a physical world view. However, common for all libertarian theories is that they are not compatible with determinism, and Morse assumes that determinism, "or something quite like it" is true. Therefore, no form of libertarian free will has a place in Morse's view of responsibility.

In Morse's view, then, the justification of the current legal system, including its retributive practices, differs drastically from the view presented in the Revision Argument. Even though Morse agrees with the revisionists (i.e., those who defend the Revision Argument) that the law is built on a folk psychological view of human action, he claims in contrast to the revisionists that the folk psychological view is not libertarian about free will. Morse holds, further, that the current legal system, including its retributive practices, is consistent with determinism and can be justified on compatibilist grounds. That is, he questions premises (2a) and (2e) in the Revision Argument. Morse points to the fact that it is entirely consistent with determinism that people can act freely in the sense of doing what they choose to do based on their reasons for action and doing so without coercion. He holds that this is the relevant form of free will and that it is sufficient for the justification of retributivist punishment.[22]

[22] Morse is far from alone in defending such a compatibilist approach to free will and retributivism. Among others, for example Nicole Vincent (2009a, 2013) Alva Noë (2010) and Michael Gazzaniga (2006) have argued in a similar manner.

In the following sections, I will discuss Morse's argument as divided into three inter-related, but distinct claims:

> (1) Legal responsibility is consistent with both compatibilism and libertarianism about free will, meaning that there is nothing in the law, on an explicit level, that requires libertarian free will in order for a defendant to be held legally responsible, but neither is there anything in the law that is incoherent with a libertarian free will.
>
> (2) The law's view of a person is a folk psychological view. Libertarian free will is not required for the folk psychological notion of responsibility. What people in general have in mind when talking about free will as a requirement for responsibility is that a person, in order to be responsible, must be able to act freely in the sense that she has the ability to choose how to act on the basis of her reasons.
>
> (3) The kind of free will that a compatibilist view of free will provides can justify legal retributive punishment.

In section 3.2, I discuss the claim Morse provides to the effect that, as the law is stated, legal responsibility is completely consistent with both compatibilism and libertarianism about free will, and some objections that have been put forward against this claim. I argue that even though these objections do not provide conclusive arguments against Morse's view that the law is neutral with regard to compatibilism and incompatibilism, it is clear that Morse's account about the law requires can be contested.

In section 3.3, I discuss two different understandings of what Morse might have in mind when he maintains that the law reflects a folk psychological view of people and their behavior. I will conclude that if Morse's position provides a substantial objection to the Revision Argument, it is most reasonable to understand his use of folk psychology as having the same explanatory role as it is taken to have in the Revision Argument, i.e., that it picks out the ordinary person's common sense view of e.g., action explanations. After this discussion, some experimental studies regarding people's view of free will are considered. These experimental studies come to different conclusions concerning the question of whether people are libertarians or compatibilists about free will and responsibility. I

conclude that there are no univocal answers to the questions whether folk psychology and folk morality are compatibilist or libertarian regarding free will and responsibility, and no clear answers to what is required for desert according to folk psychology and folk morality, either. Hence, we cannot draw any straightforward conclusions as to whether it is true or not that folk psychology is compatibilist regarding free will. (However, in chapter 6, I will return to this question and argue that, at least on one construal of folk psychology and folk morality, it is probably not compatibilist.)

In section 3.4, some challenges to Morse's claim that free will compatibilism can provide a secure basis for retributive punishment is spelled out. I will argue that if moral responsibility (in the sense of retributive desert moral responsibility) is based on a certain mental capacity (in other words, if a certain mental capacity is what makes someone morally responsible), this capacity must be relevantly different from capacities that do not make someone morally responsible. This argument departs from what will be called "the Principle of Relevant Difference" which is spelled out in 3.4.1.[23]

What can be picked out as a relevant difference is, I argue, restricted by one's metaphysical commitments. In the current context, the two most significant constraints for what can be referred to as a relevant difference are physicalism and determinism.[24] In short, for any natural property (as for example, a mental capacity) that is claimed to be the relevant difference between a person who is morally responsible and a person who is not, this natural property cannot involve features of indeterminism, and it must be possible to explain within a physicalist explanatory framework.

---

[23] The Principle of Relevant Difference is intended to be a general principle, and it does not only apply to compatibilism. For example, according to libertarianism about free will, the relevant difference with regard to moral responsibility is the presence or absence of libertarian free will. But since libertarian free will is not compatible with determinism, this is not an alternative for the discussions in this thesis.

[24] As I have already pointed out, physicalism in a liberal form (i.e., not reductive physicalism) is embraced by nearly all free will philosophers. Determinism is a more controversial metaphysical thesis. The reasons for why it is accepted as a premise in the discussions of this book are described in chapter 1, where I also shortly discuss some different accounts of libertarian free will.

3.2 Legal responsibility and compatibilism

According to Morse, the law is consistent with *both* compatibilism and incompatibilism about free will, since there are no explicit references to any of these metaphysical positions in criminal law doctrine. The criteria for legal responsibility and culpability rely on neither compatibilism nor libertarianism, as a foundation for their justification.

> […] it is crucial to recognize that libertarian free will is not an element of any crime or of any affirmative defense. To establish prima facie guilt, the prosecution *never* needs to prove that the defendant had free will. To establish an affirmative defense, the party with the burden if persuasion *never* needs to prove the presence or absence of free will…in short, free will or lack of it is not a criterion for criminal responsibility or non-responsibility. Once again, it is irrelevant to the actual practices of criminal law. (Morse, 2013, p.38)

Morse states that what is crucial for criminal responsibility is not free will but instead several other elements:[25]

> There are five types of elements that define crimes: acts (sometimes referred to as 'conduct' elements), an accompanying mental state (termed *mens rea)*, attendant circumstances, results, and causation (in cases in which there is a result element). (2013, p.35)

Morse provides us with an example of a crime that involves all five elements: "the intentional killing of a police officer knowing that the victim is a police officer" (2013, p.35). The act element is required for someone to be held criminally responsible. In the example, the act is intentional killing. The accompanying mental state, *mens rea*, picks out the intention that the agent had at the time of action. In the example, the intention was to kill. If the agent would instead have had the intention to scare the victim, or perhaps had acted in self-defense, then the agent would have lacked the intention to kill. In Morse's

[25] The claim that free will is not included as an element that defines criminal actions, and that free will is not needed in order to establish prima facie guilt and an affermative defense is probably embraced by most legal theorists. In this regard, Morse's position does not challenge the common view of these cases. However, as we will see, not everyone agree that this fact implies that free will is irrelevant to the assessment of legal responsibility.

example, the agent did not only have the intention to kill, but also knew that the victim was a police officer, which makes it a more specific crime. The result element is the death of the victim – if the victim would not have died, the defendant could not be guilty of more than an attempt to kill. Finally, the victim's death must be the consequence of the defendant's action. If the victim dies for any other reason, the "so-called causal chain between the defendant's conduct and the victim's death might be 'cut'" (2013, p. 37).

As Morse describes it, what matters when assessing if someone deserves legal punishment is what was going on in the defendant's mind at the time of action, but also several external circumstances such as what actually took place at the time of the act regardless of the defendant's intentions. However, none of these criteria, neither the "subjective" nor "objective" ones, are incompatible with a deterministic world-view. The same holds for the excuses. A person who has done something wrong can be excused if she e.g., has a general incapacity of acting rationally, or were temporarily incapable of rational behavior at the time of action. These facts, according to Morse, "explains why young children and some people with mental disorders are not held responsible" (2015, p. 257). Furthermore, compulsions and coercion can also be excusing conditions under certain circumstances, since these conditions defeat the requirement of a voluntary act. It can be that someone suffers from a disease that results in uncontrolled bodily movements, or that someone is coerced (by an external agent) to do something unlawful. But there are also excusing conditions linked to internal compulsive states that are defined as mental disorders, which often involves loss of action-control, as e.g., auditory hallucinations that command a psychotic patient to commit a crime. The reason for why the presence of certain mental states may provide an excuse has nothing to do with determinism, but is connected to the agent's ability to make rational choices: "All of the distinctions that criminal responsibility criteria draw are consistent with retributive and consequential theories of just blame and punishment that we endorse and with the truth of determinism" (2015, p. 258).

Morse does not only claim that free will is irrelevant to law in the sense described above, i.e., that the criteria for legal responsibility can

be accounted for in a manner that is consistent with the truth of determinism. He also recommends forensic psychiatrists and psychologists to avoid thinking and talking about free will in their work, since free will is not an issue in forensic assessments:

> Forensic psychiatrists and psychologists should void all mention of free will in their reports, testimony and scholarship. They should not even think about free will as an issue in forensic work. Using the concept of free will can only confuse oneself and the legal agents to whom our work is addressed. It can never properly be a premise or conclusion in any forensic argument. It can never clarify any legal issue or help resolve any legal case. If one has a taste for deep philosophical problems, free will is of course worth thinking about. The issue is an endlessly interesting evergreen that will never be solved to everyone's satisfaction. But if one thinks about the problem in this sense, one is doing philosophy, not forensic work. Some people think philosophy is a disease, however, so be forewarned. (Morse, 2007, p. 220)

However, there are scholars from various fields that disagree with Morse on this account. Meynen writes:

> Morse's argument is made within the context of the U.S. legal system. At least in principle, other systems may mention freedom of will in their legal doctrines, documents, or other relevant legal sources regarding insanity. For instance, in the Netherlands, "free will" is mentioned in some verdicts where the court explains why the defendant is not criminally responsible. (Meynen, 2016, p. 68)

Moreover, Meynen maintains that even if Morse were right about the fact that free will is not mentioned in legal doctrine, this does not necessarily entail that free will is irrelevant to the law. As was discussed in section 2.5, responsibility and free will are regularly assumed to be closely connected, and according to many scholars, acting out of free will is just to satisfy the requirements for being responsible for one's actions (O'Connor & Franklin, 2019). According to a view where free will and responsibility are closely connected, we do not need to explicitly refer to free will when attributing responsibility, since what we mean by saying that someone is responsible is, partly, to say that she has acted out of free will.

Other thinkers have argued that the U.S. legal system itself is not as neutral regarding free will as Morse claims it to be. Philosopher and neuroscientist Sam Harris writes:

> The U.S. Supreme Court has called free will a "universal and persistent" foundation for our system of law, distinct from "a deterministic view of human conduct that is inconsistent with the underlying precepts of our criminal justice system (*United States v. Grayson* 1978). Any intellectual developments that threatened free will would seem to put the ethics of punishing people for their bad behavior in question. (Harris, 2012, p. 48)

In a similar vein, Michael Moore argues that hard determinism provides a challenge to our moral and legal practices, in the sense that we must be able to answer to the questions posed by determinism – as for example, how to make sense of the idea that agents control their own actions – in order for the system of responsibility attribution to make sense. Moore writes that we *implicitly presuppose* rather than *explicitly defend* answers to why certain elements are relevant when distinguishing the excused from the responsible (Moore, 2016, p. 48). These presuppositions about how we distinguish between the culpable and the excused are crucial to the whole system of responsibility attribution since this system only makes sense in light of such presuppositions:

> [T]he internal, work-week job of distinguishing the excused from the responsible can only be done if one has ready-to-hand certain answers, but not others, to the big, external questions posed by hard determinism. It may be that we who make daily judgements of responsibility and excuse can implicitly presuppose rather than explicitly defend those answers to hard determinism; but we cannot do our internal work without relying on such answers. It is not just that the whole system of responsibility attribution makes sense only if such answers can be defended. This is true enough, but also true is the fact that internal judgements about particular excuses will take their discrete shape in light of such general answers to hard determinism. We "internal" moralists and criminal lawyers have to go external to do our jobs. (Moore, 2016, p. 48)

Moore's view stands in sharp contrast to that of Morse's, who claims that what Moore calls "external" questions are generally irrelevant for legal responsibility assessment.

In sum, Morse's statement that questions about free will and determinism are not an issue when it comes to legal responsibility, and that the law is "silent" when it comes to questions about compatibilism and incompatibilism has been questioned by scholars in the field. As both Meynen and Harris point out, it is not obvious that the legal practice is as silent when it comes to free will as Morse claims it to be, and Moore argues that forensic psychiatrists and lawyers have to go "external" to be able to answer fundamental questions concerning their "internal" everyday practice. But as a matter of fact, Morse do discuss external questions in a sense, since he claims that the folk psychological conception of free will is fully compatible with determinism, and it is folk psychology that legitimizes the law. So even though the law as such is "silent" about free will, legal practices must still be compatible with the intuitions regarding free will, responsibility and punishment that are present in folk psychology (and folk morality). This means that if we accept Morse's claim that the law is "silent" about free will, this is not a definite defense against the claim made in the Revision Argument: that legal retributive punishment presupposes libertarian free will. Since what the Revision Argument is ultimately concerned with is the justification of the "internal" practices of the law, not what is explicitly stated in the law. It seems that Morse agrees with the claim made in the Revision Argument: that the internal practices are justified by folk psychology and folk morality.

So, *if* legal retributive punishment, according to folk psychology, requires that people have a libertarian free will in order to be justified, *then* the Revision Argument stands and it does not really matter if the law in itself is "silent" or not about free will. But if legal retributive punishment does not, according to folk psychology, require libertarian free will, then the Revision Argument is built upon a faulty premise.

Morse, meanwhile, argues that folk psychology is compatibilist about free will and responsibility, i.e., that legal retributive punishment does not require libertarian free will for its justification.

In the following sections, this claim is scrutinized. First, I will say a few words about what folk psychology is, since the meaning of this concept is not straightforward. Second, I will discuss some results from experimental philosophy carried out by different research groups to try to highlight how "ordinary people" actually do think about free will, responsibility and punishment.

3.3 Folk psychology, folk morality and the justification of retributive punishment

3.3.1 Folk psychology

Folk psychology remains a vague concept. In the present context, folk psychology is used to describe a common and pre-theoretical way of thinking of, predicting, and explaining human behavior. In other words, when "we" are thinking of, predicting, and explaining human behavior in terms of what people have in mind, e.g., in terms of mental states, we normally do this within the folk psychological framework.

According to eliminative materialism, folk psychological notions characterizing e.g., mental states can be replaced with neuroscientific descriptions of what is going on in the brain when someone has, for example, a belief. Two of the most well-known proponents of eliminative reductionism are P.M Churchland (1981, 1989) and P.S Churchland (1986) who both argue that folk psychology is an inaccurate way to understand human thinking and behavior. Within folk psychology, human behavior is described and explained as it appears to us in an everyday context. But, the Churchlands argue, as we in the last decades have gained a lot of knowledge about how our brain works, folk psychology appears like an outdated way to explain behavior and we should replace folk psychological concepts and explanations with neuroscientific concepts and explanations. The latter will provide us with a more accurate understanding of why people behave as they do.

As previously mentioned, (in section 1.6), physicalism does not necessarily entail that all kinds of phenomena can be reduced to physical phenomena. Non-reductionism about mental states can, for

example, be defended on the grounds that mental concepts are such that they cannot be "translated" into, or fully analyzed in terms of, concepts that refer to the brain. Certain predicates are essential for a full description of the world, and they are not reducible to physicalist predicates. Mental predicates cannot, according to this view, be reduced to physical predicates, because mental states are classified in functional terms rather than natural kind terms (Robinson, 2017). (I will return to this view of mental terms and the implication it may have for our responsibility practice in chapter 5.)

Let us return to Morse: he agrees with the claim made in the Revision Argument to the effect that the current legal system is based on a folk psychological understanding of human behavior. According to eliminative reductionists, like the Churchlands, this is potentially problematic since folk psychology is an *inaccurate* way to describe human behavior, in so far as we do not want our legal system to rely on inaccurate descriptions and explanations. If we take the non-reductionist perspective, it need not be a problem that the legal system is based on a folk psychological view of behavior, since folk psychology may be an equally good (or better) way of explaining behavior as any other way, as for example a more scientifically minded approach.

Greene & Cohen (2004) suggest there may be a middle ground between eliminative reductionism and non-reductionism when it comes to folk psychology and the law. It may not be needed to eliminate the folk psychological terminology from our language, just because it is inaccurate. We should, however, recognize that it has certain problematic features, and therefore use a better perspective– the scientific one – when constructing legal standards. The reason why we should make a difference between our day-to-day judgments and judgments that guide how our laws should be constructed is that the law must live up to higher standards with regard to moral justification, than our everyday moral judgments need to. For example; if criminal law is justified in sentencing certain people to prison for a long time for their crimes, this moral practice must be robustly justified, both morally and epistemically (for instance, with proof beyond a reasonable doubt.) Even if we legitimately ignore certain counter-intuitive truths about human behavior in our everyday

life, we cannot legitimately ignore it in legal contexts where the stakes and thus standards for moral and epistemic justification need to be much higher (Greene & Cohen, 2004, p. 1784).

Greene & Cohen note that there is a fair amount of evidence suggesting that humans have a set of specialized cognitive subsystems both for processing information about intentional agents, but also for how ordinary matter behaves. A central feature in how we understand intentional agents is, they claim, that agents are uncaused causers: we need thus to understand why agents move around not in terms of apparent physical causes, but rather because of features of mind such as beliefs, desires and intentions. It is in relation to these features that ideas of free will, praise and blame come into the picture. Referring to Daniel Wegner (2002), Greene and Cohen argue that since we experience ourselves as being uncaused causers, and consequently feel as if we, in contrast to physical objects, cause our own actions, we "imagine that we are metaphysically special, that we are non-physical causes of physical events" (p. 1781).[26] And since we place ourselves in the category of "intentional agents", it is natural for us to explain other people's behavior as free as well. So far, Greene & Cohen's account is compatible with eliminative reductionism. However, they do not argue that we should *replace* the folk psychological account of human action – because they doubt that it is a possibility open to us:

> After thousands of years of our thinking of one another as uncaused causers, science comes along and tells us that there is no such thing, that all causes, with the possible exception of Big Bang, are caused causes (determinism). This creates a problem. When we look at people as physical systems, we cannot see them as any more blameworthy or praiseworthy than bricks. But when we perceive people using our intuitive, folk psychology we cannot avoid attributing moral praise and blame... The problem of free will and determinism will never find an intuitively satisfying solution because it arises out of a conflict between two distinct cognitive subsystems that speak different cognitive 'languages' and that may ultimately be incapable of negotiation. (Greene & Cohen, 2004, pp. 1782-83.)

[26] Smilansky (2000) argues for the same conclusion in *Free Will and Illusion,* p. 26, as well as Strawson (1986) in *Freedom and Belief,* pp. 281-84.

Greene & Cohen suggest that it may be psychologically impossible to view ourselves and others from any other perspective than the intuitive, folk psychological one which comes together with assumptions about free will, blame- and praiseworthiness, etcetera. But when we decide the legal requisites and procedures for when it is morally permitted to punish people, we should apply a scientific perspective. Even though it is practically impossible for us to live by the fact that we have no free will, free will is still an illusion. Evolution has made us experience ourselves in this illusionary way, but we should not mistake this illusion for being real. Science makes us realize that. Our folk psychological view of behavior will take us astray when it comes to metaphysical questions of why people behave as they do, and we need to have a correct – evidence-based – view of these matters when we form our legal system. For example, we should not form our legal system based on the idea that people have a libertarian free will, and that they can deserve punishment because of that. Instead, legal practices should be based on e.g., consequentialist considerations.

As the above section shows, Greene & Cohen make some substantial claims about the folk psychological view: they move from an explanation of why we feel as if we are free, uncaused causers, to explanations of how we transfer these experiences into metaphysical beliefs, and further to the argument for why this metaphysical view should not be relied on when constructing legal standards. It is this account of folk psychology that they refer to when they claim that the law's view of the person is folk psychological, and it is this substantial account of folk psychology they argue we need to abandon when it comes to discussions of justified legal punishment.

Compared to Greene & Cohen, Morse seems to apply a much more open (in the sense of less detailed) conception of folk psychology in his argument:

> Folk psychology does not presuppose the truth of free will, it is perfectly consistent with the truth of determinism, it does not hold that we have minds that are independent of our bodies (although it, in ordinary speech, sound that way), and it presupposes no particular moral or political view... the definition of folk psychology being used here does not depend on any particular bit of folk wisdom about how

> people are motivated, feel, or act [...] The definition of folk psychology presupposes and insists only that human action can at least be partially explained by mental state explanations or that it will be responsive to reasons, including incentives, under the right conditions. (Morse, 2013, p. 31)

The conception of folk psychology that Morse has in mind "does not depend on any particular bit of folk wisdom about how people are motivated, feel, or act." What our "folk wisdom" comes down to remains a bit unclear in Morse's writings. Folk psychology does not entail any special metaphysical view of mental states, beside the idea that mental states are central in explanations of action and that people are, in general, responsive to reasons: "the definition of folk psychology presupposes only that human action can at least partially be explained by mental state explanations or that it will be responsive to reason, including incentives, under the right conditions". Apparently, Morse has a different conception of folk psychology in mind when he claims that the law's view of the person is a folk psychological view, compared to what Greene and Cohen have in mind when they make similar claims.

If we for a moment return to the Revision Argument, we need to remind ourselves of why folk psychology features in the discussion in the first place. In the Revision Argument, it is claimed that the law's justification of retributive punishment is grounded in a folk psychological (and folk moral) view of behavior, and more specifically the folk psychological assumption that people have libertarian free will (2a), and that this is required in order to deserve punishment (2e). According to Morse's counterargument, even though the law is indeed founded in a folk psychological view of behavior, folk psychology is not necessarily libertarian. Still, it presupposes that actions can be explained with reference to mental states, and also that people are reason responsive.

It is reasonable to think, as Morse does, that folk psychological explanations of behavior do not *always* have to include allusions to free will. We can often explain behavior with reference to mental states without making any statements, or having any specific beliefs, about whether a person acted out of free will or not. For example, a natural explanation to why I just went to the lunch room would be

that I was hungry, and believed that there was an apple left in the fruit basket. No references to free will (explicit or implicit) are necessary in order to understand how my mental states explain my behavior. However, the question at stake here is whether the Revision Argument is correct in that a libertarian free will, according to folk psychology and folk morality, is required for the justification of retributive punishment.

There are at least two interpretations of Morse's claim about folk psychology: First, Morse uses a different concept of folk psychology than the one used in the Revision Argument. And according to his stipulated concept of folk psychology, libertarian free will is not included. Since the law's view of a person is the folk psychological view, libertarian free will is not a question for the law, either. This claim would be true in virtue of the stipulated definition. The first interpretation is suggested by Morse's talk of "the definition of folk psychology being used here," as he puts it in the quote above. Second, Morse claims that "folk psychology" refers roughly to how people actually, regardless of stipulated definitions, think of and explain human behavior. In this sense, Morse's claim, as well as the rivalling claim about folk psychology put forth in the Revision Argument, can be contested on empirical grounds: if we understand folk psychology as "the way people commonly think of and explain human action," what is presupposed in a folk psychological explanation is an empirical question.

Since the Revision Argument states that the legal system is construed upon folk psychological understanding of behavior and legitimized by folk morality, what Morse says about folk psychology is only an objection to the Revision Argument if Morse talks about it roughly in the same sense as it is used in the Revision Argument, namely, that "folk psychology" refers to ordinary people's common sense intuitions of why people behave as they do, and what considerations that, according to them, can be brought into an explanation of actions. Moreover, folk psychology is the framework

within which folk morality operates, that means, moral judgments are connected to folk psychological explanations.[27]

In the following sections, I will discuss Morse's objection to the Revision Argument according to this interpretation: Morse disagrees with Greene & Cohen about the ordinary person's common sense understanding of how actions are explained and what is required for deserving punishment.

If the content of folk psychology is an empirical question, it is important – if not crucial – to enquire about the actual beliefs ordinary people have regarding the topics under investigation. The next section, therefore, is concerned with some experimental studies of people's moral intuitions concerning free will, responsibility, determinism and desert.

3.3.2 Folk morality & experimental philosophy

What do "ordinary people" actually think about the compatibility of determinism and free will? According to Nichols (2015, p. 31) experimental results show that there is a cross-cultural tendency among people to be incompatibilists and indeterminists about their own choices when it comes to explaining their ability to do otherwise. Moreover, Knobe & Nichols (2007) suggest that when "ordinary people" appear to be compatibilists, this is often because of performance errors. They present evidence that 86% of the participants in their study judged that it is not possible for an agent to be "fully morally responsible" in an abstract, deterministic scenario that does not specify a particular agent or action. However, when the participants are presented with a concrete case that engages people's emotions, 72% of the participants judged that Bill who coldheartedly

[27] Neither Morse nor Greene & Cohen explicitly discuss folk morality as distinguished from folk psychology. In their writings, it seems as they use "folk psychology" to cover what I mean with both "folk psychology" and "folk morality." Folk morality and folk psychology are intertwined in these discussions, since folk morality is based on a folk psychological understanding of how people think and act, but I think it is useful to treat them as separate phenomena since folk psychological explanation must not include folk moral elements. As will be pointed out later on in this thesis, it is also conceivable that folk morality can be based on other explanations than folk psychological ones.

killed his wife and children was "fully morally responsible." So, in the abstract scenario, the vast majority of ordinary people do not believe that full moral responsibility is possible in a deterministic world, while in the concrete scenario, the vast majority of ordinary people do feel that full responsibility is compatible with determinism. So, the experimental design (the type of cases that ordinary people have to make judgments about) may be very important for the type of experimental results one gets. Frank Jackson argues that we must be careful when assessing experimental results, and take into account how people respond to a variety of different cases in order to reveal what their real intuitions are:

> A person's first-up response as to whether something counts as a K may well need to be discounted. One or more of the theoretical role they give K-hood, evidence concerning other cases they count as instances of K, signs of confused thinking on their part, cases where classification is, on examination, a derivative one (they say it's a K because it is very obviously a J, and they think, defensibly, that any J is a K), their readiness to back off under questioning, and the like, can justify rejecting a subject's first-up classification as revealing their concept of K-hood. (Jackson, 1998, p. 35)

In line with Jackson's advice, Nichols & Knobe suggest that the best explanation for the inconsistency in intuitions found in studies is that concrete examples engage people's emotions in a way that abstract cases do not, and thereby leading people to offer apparent, or "first-up" compatibilist judgments. But since people, when not biased by strong emotions regarding the case, report incompatibilist intuitions, we could also explain away these compatibilist results as performance errors and conclude that people's underlying theory is incompatibilist (2007, p. 672). Other studies have confirmed that the intuitions regarding compatibilism and incompatibilism are sensitive to how abstract the test case appear. For example, Roskies and Nichols (2008) observed that, *ceteris paribus,* people are expressing more compatibilist intuitions towards scenarios taking place in our world rather than in an alternate universe.

However, Murray & Nahmias (2012) report, in line with Morse's claim, that they found opposite results compared to the findings reported by Nichols & Knobe. Murray and Nahmias' results suggest

that most people have compatibilist intuitions and that incompatibilist results can be explained away since they are grounded in a bad understanding of determinism. Nahmias, Coates, and Kvaran (2007) provide some evidence for the claim that people tend to report different intuitions depending on whether an agent is described as psychologically determined, as opposed to neurologically determined. Generally, cases described as neurologically determined tend to generate more incompatibilist intuitions compared to cases described as psychologically determined. Nahmias & Murray suggest that this result is explained by the fact that people (mistakenly) tend to think that neurological determinism "bypasses" mental causation – in the sense that the neurological deterministic explanations of actions entail that the agent's mental states have no causal role for her actions – whereas psychological determinism does not elicit this intuition. But, they argue, to think that determinism bypasses mental causation is a misunderstanding of determinism. Nahmias and Murray constructed a test in order to test this hypothesis. In the test, Nahmias and Murry randomly assigned the participants with four different descriptions of determinism. The participants were then asked to indicate their level of agreement to a series of statements on a six-point scale, ranging from "strongly disagree" to "strongly agree." The questions concerned not only whether the agent in the described scenarios was fully morally responsible, can have free will, and deserved praise or blame for their actions, but also whether the agent's desires, beliefs, and decisions have an effect on what the agent ends up doing, and whether the agent has control over what he or she does (pp. 445- 46). According to Nahmias & Murray, analyses of the data strongly confirmed the hypothesis that there is a correlation between beliefs that determinism bypasses mental causation on the one hand and incompatibilist intuitions on the other.

According to Nahmias & Murray these results mean, in contrast to the conclusion from Nichols & Knobe, that it is the incompatibilist intuitions that are grounded in a performance error, an error due to a misunderstanding of the relation between determinism and mental causation. Moreover, they refer to a test constructed to control for this result, which shows that the incompatibilist intuitions indeed decrease when the case presented explicitly mentions that

determinism does *not* mean that people's mental states (their beliefs, desires, and decisions) have no effect on that they end up doing, and it does *not* imply that people have no control over their actions.[28]

The conclusions that can be drawn from the experimental research described above is that there is some evidence pointing in the direction that people in general have compatibilist intuitions when it comes to free will, responsibility and determinism, that supports Morse's view, but there is also some evidence that points in the other direction, i.e., that people have incompatibilist intuitions. Experimental philosophy does not provide us with a straightforward answer to the question of whether ordinary people have compatibilist intuitions or not.[29] The question of how to understand folk morality and its connection to folk psychology and science will be further discussed in chapter six.

---

[28] To my mind, the interpretation Nahmias & Murray provide for their test results is not obviously the most plausible one. Their test results can be understood in at least two different ways. Either, we can understand it, as Nahmias & Murray do, as that people who report incompatibilist intuitions conflate determinism with bypassing of mental causation. This means that people make the mistake of assuming that if determinism holds, mental states are not causally efficacious. This would, of course, be a mistake: determinism as such doesn't make mental causation impossible. Accepting determinism only means that mental causation would work in accordance with determined laws, in contrast to mental causation as, e.g., an 'uncaused causation' phenomenon. But there is another possible interpretation that is compatible with the results that Nahmias & Murray present. According to this interpretation, what determinism bypasses is not mental causation in the metaphysical sense, but the significance, or explanatory power, of mental causation, with regard to free will, responsibility and desert. This line of thought relates to what Björnsson & Persson call the 'Explanation Hypothesis' (Björnsson & Persson, 2011, p. 4) According to Björnsson & Persson, an agent is assessed as responsible for an event when the agent's motivational structures are "a significant part of the explanation of such an event" (Björnsson & Persson, 2011, p. 3). However, for the purposes of the current discussion, it suffices to conclude that experimental philosophy provides us with unsatisfactory results regarding the question of whether 'common people' are compatibilists or incompatibilists about free will and responsibility.

[29] There are interesting experiments suggesting that individual differences with regard to people's character traits play a role in how prone they are to ascribe responsibility in a deterministic world. Feltz & Cokely (2009) found, for example, that people high in extraversion were more likely to judge an agent in a deterministic world as free and responsible compared to more introvert people.

This means that Morse's claim that folk psychology is fully compatible with determinism cannot be verified on empirical grounds (neither can it be dismissed). However, Morse does not only rely on the claim that folk psychology is compatible with determinism for his conclusion that retributive legal punishment is not threatened by the Revision Argument. He also refers to the philosophical discussion concerning determinism, responsibility and free will, and argues that free will compatibilism "[...] provides the only secure basis for criminal responsibility" (Morse, 2013b, p. 27). He continues: "I do not aim to argue for the truth of compatibilism. It is sufficient that it is one of the two plausible positions in the metaphysical debate – the other being hard determinist incompatibilism – and, in one form or another, it is probably the position held by the vast majority of professional philosophers" (Morse, 2013b, p. 28)

Morse is probably right about that compatibilism, in one form or another, is the most popular position among contemporary (analytical) philosophers when it comes to responsibility and free will. However, in the current discussion we are specifically concerned with the fact that basic desert is required in order for retributive punishment to be justified. And it is not necessarily the case that compatibilism about free will and responsibility entails compatibilism about basic desert. In other words, it is not obvious that philosophers that accept compatibilism about free will and responsibility also accepts that the kind of free will responsibility they advocate allows for basic desert attribution. As e.g., Caruso & Morris (2017) point out, it is not always clear that philosophers, when they discuss moral responsibility, have "retributive desert moral responsibility" in mind, they can very well think of a notion of moral responsibility that allows for punishment on other grounds than basic desert.

But, as I understand Morse, he holds that basic desert compatibilism is more or less an entailment of free will compatibilism. I will point to some challenges for this view in the following section.

3.4 Challenges for compatibilist basic desert retributivism

3.4.1 The Principle of Relevant Difference

In order to scrutinize Morse's claim that free will compatibilism can justify legal retributive punishment I will discuss some common compatibilist theories and whether they can provide what is required for basic desert. However, we must be aware of what we are searching for in this discussion – what is it, more specifically, that a compatibilist theory must provide in order to justify basic desert attribution? In this section I will introduce a principle that I will call "the Principle of Relevant Difference." I will argue that if a compatibilist theory can account for moral responsibility, it must meet the demands of this principle. It goes like this:

> THE PRINCIPLE OF RELEVANT DIFFERENCE
>
> When two phenomena (actions, agents, events, capacities, abilities, etc.) differ with regard to moral properties, the difference in moral properties must be due to a relevant, non-moral difference between those two phenomena. For example, if one action is wrong and another is not, this difference in moral status must be due to a relevant, non-moral difference between the two actions.

This principle is related to (but not identical with) the idea that there cannot be an ethical difference between two states of affairs or actions without there being some natural difference between them. That there is a necessary connection between natural and moral properties, was labeled "supervenience theses" by R.M. Hare (1952), and the view that moral properties supervene on natural properties is widely acknowledged by moral philosophers. In Henry Sidgwick's *The Methods of Ethics,* he writes:

> In the variety of coexistent physical facts we find an accidental or arbitrary element in which we have to acquiesce...But within the range of our cognitions of right and wrong, it will be generally agreed that we cannot admit a similar unexplained variation. (Sidgwick, 1907, p. 209)

In the contemporary meta-ethical discussion, Michael Smith (2004) expresses a similar idea:

> Virtually everyone writing in meta-ethics takes it for granted that evaluative facts supervene on natural facts [...] I trace the attraction of the supervenience thesis to a fact about ordinary moral discourse, namely, the fact that it is always appropriate to ask what makes a moral claim true and what we require by way of a response is an answer in terms of certain natural features. (Smith, 2004, p. 10)

Ethical supervenience is, hence, a relatively uncontroversial idea. Tristram McPherson describes the basic content of this thesis as follows:

> Suppose that a bank manager wrongfully embezzles his client's money. If we imagine holding fixed how much the bank manager stole, and how; the trust his customers placed in him; what he did with the money; all of the short- and long-term consequences of his actions; and so on, it seems that there could not be a second action that perfectly resembled this embezzlement, except that the second action was right rather than wrong. Cases like this one seem to show a *necessary* connection: they suggest that the ethical character of the bank manager's act cannot vary without some other facts varying as well. (McPherson, 2015)

McPherson writes that "the ethical character of the bank manager's act cannot vary without some other facts co-varying as well." McPherson then notes that the most common way to characterize ethical supervenience in the literature is in terms of moral properties supervening on *natural* properties, even though there are some difficulties with this idea. For example, the term "natural" lacks a precise and canonical definition. I will put such difficulties aside here, and in what follows, I will discuss ethical supervenience as a supervenience relation between ethical properties and natural properties.

According to the supervenience thesis, two actions cannot have different moral properties (or, in other words, different ethical character) without a difference in natural properties. However, in my view, it is intuitively plausible to reinforce the ethical supervenience thesis without making it much more controversial: according to this reinforced supervenience thesis, the claim is that for every difference

in ethical character, this difference is due to a natural difference *of a relevant sort*. Not just any difference will do the job. The Principle of Relevant Difference captures this point. What a relevant difference consists in will be discussed further in the next subsection.

There is another supervenience relation that will play a central part in the discussions about whether compatibilism can meet the demand from the Principle of Relevant Difference: that between mental states and brain states. Here I will basically follow Davidson's characterization when I talk about how mental states supervene on physical (brain) states. Davidson characterizes the supervenience relation between the mental and the physical as follows:

> [M]ental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience 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. (Davidson, 1970, p. 214).

Supervenience claims are, in themselves, not particularly explanatory forceful. For example, the claim that mental property A supervenes on physical property B is just a claim about a certain pattern of property co-variation: it does not tell us why this pattern holds, or the nature of the dependency (Kim, 1993, p. 167). Even though these explanatory questions are pressing if we want to understand *why* mental states co-vary with physical states, this is not the primary concern in the following discussions. For the purposes of the discussion here, it suffices to say that when a specific property, or set of properties, as e.g., mental states, supervene on another property, or a set of properties, as e.g., physical states, this means that the property that supervenes (i.e., the supervenient property) co-varies with the property it supervenes upon (i.e., the base-property) and that this is not an accidental co-variation, because every instance of the supervenient property will co-vary with the supervenience-base and it cannot be otherwise.

Kim suggests that there are two different concepts of supervenience, one with a stronger modal force than the other (Kim, 1984, p. 157). *Weak* supervenience is, according to Kim, when the co-

variation between supervenient properties (e.g., the mental properties) and base-properties (e.g., the physical properties) holds *within* possible worlds, but not *between* possible worlds. In other words, the fact that there is a specific co-variation between a set of base properties and a set of supervenient-properties in a world W_1 does not entail that this co-variation must necessarily hold in world W2. In W_2, the supervenient properties may co-vary with another set of base-properties (pp. 159-169). Strong supervenience is, according to Kim, when the co-variation between the supervenient properties and the base-properties holds *necessarily*, i.e., that for every two possible worlds such that people are indiscernible with regard to the base-properties, they are also indiscernible with regard to the supervenient properties (p. 165).[30]

If we return to moral supervenience, Kim notes that weak supervenience falls short of accounting for the intuitive sense in which moral properties supervene on natural properties, since this intuition is often taken to imply that any two worlds exactly alike in all natural respects must be alike in all moral respects. However, weak supervenience still meets the "consistency requirement," which is based on a principle that Kim calls "the principle of universalizability" and which, in turn, is based on the intuition that ethical judgments should be generalizable, i.e., the idea that like cases should be treated alike (pp. 161-162). For the discussions in this thesis, weak supervenience will do the job, since what we are interested in is the similarities and differences between cases that are treated differently with regard to basic desert within the same world. Retributivist legal systems need to distinguish between people who deserve to be punished for their actions, and those who do not – in the actual world. And we are interested in the question of whether these (actual world) cases are sufficiently different with regard to natural properties in order to differ with regard to the moral property of basic desert. If this is not the case – i.e., if we cannot find a sufficient difference in

[30] "Necessarily" can be interpreted in different ways. For example, there can be logical, metaphysical and nomological necessity. According to Kim, how to specify "necessarily" depends on the particular supervenience thesis under consideration: "We should […] leave an exact interpretation of 'necessarily' as a parameter to be fixed for particular cases of application" (Kim, 1984, p. 166).

the natural properties (i.e., the base properties) – then to ascribe different moral properties to them would arguably violate the consistency requirement.

3.4.2 What is a relevant difference?

A basic assumption behind the Principle of Relevant Difference is that moral properties supervene on natural properties. If two people have identical natural properties (i.e., identical base-properties) they will be identical with regard to moral properties (i.e., they will have identical supervenient properties). In this sense, the supervenience idea is the very plausible idea that we should "treat *completely identical* cases alike." But it also seems plausible that we should "treat *sufficiently similar* cases alike." For example, when two people share the moral property of "being a good person," and this moral property supervenes on their natural properties of being caring and loving people, it does not have to be the case that they are *exactly* alike in these aspects in order for both being good. Even if one of these two persons is a bit more loving than the other this difference is not enough to make one of these two people good and the other not good. They are still *sufficiently* similar in order to be "treated alike."[31] This point is what the relevant difference condition in the Principle of Relevant Difference is about. The principle holds that if two things are sufficiently and relevantly similar with regard to their natural properties, they will also be alike with respect to their moral properties. So, in order for two people to differ with regard to moral properties, this difference must be due to a difference in natural properties *of a relevant kind*. Just *any* difference between natural properties cannot do the job of explaining a difference in moral properties. If a person A has a certain natural property and person B lacks this natural property, this does not necessarily entail a difference between A and B with regard to a certain moral property.

Even if the above reasoning says something over and above the supervenience thesis as it is spelled out by e.g. Hare, I take it to be a fairly unproblematic view. The idea can be illustrated if we consider

[31] The meaning of "sufficiently similar" cannot be given a detailed description, but is most plausibly assessed on a case-to-case basis.

McPherson's example that was mentioned in section 3.4.1, in which a bank manager wrongfully embezzles his client's money. Consider a situation in another world in which the situation is exactly like McPherson's example: the bank managers in these two worlds are two exactly identical copies of one another, physically, psychologically, historically, and so on, and the worlds are also almost identical. The only difference between them is that one of the bank managers has green eyes and the other has brown eyes. But this difference in natural properties between the bank managers seems, intuitively, not relevant for making a moral difference between them with regard to the question if they deserve punishment for having embezzled money from their customers. With regard to the *relevant natural properties* that determine the ethical character of the actions and the bank managers themselves, there is no relevant difference. Obviously (and non-surprisingly) it is intuitively implausible that the ethical character of an action supervenes on *any* kind of natural property. Instead, some natural properties are relevant to moral properties and others are not. In order to determine if two states of affairs, actions, or persons differ concerning moral properties, it does not suffice to identify that there is a natural difference of any kind between them. We must be able to identify whether the natural properties that differentiate between two states of affairs, actions, or persons *can account for a difference in moral properties*.

A difference that intuitively is relevant, in contrast to eye color, is behavior. In an alternative world where none of the bank managers embezzled any money, but one of them always care for his colleagues and customers, whereas the other only cares about his own interests, there seem to be a difference between *relevant* natural properties with regard to the moral property of being a good person. Thus, this difference in natural properties may be able to account for why one of them has the moral property of being a good person while the other lacks this moral property. Their actions can be compared similarly: two actions differ with regard to their moral properties only if they differ in a *relevant sense* at the level of natural properties. Let us summarize: to say that having/lacking the natural property G is a relevant difference with regard to having/ lacking the moral property F (e.g., the moral wrongness of an act) is to say that the fact that one

act has property G and another act lacks property G can make it the case that one act has F and the other does not (e.g., that one act is wrong and the other is not). In other words, it is to say that G is plausibly a natural property that can make an action wrong.

Morse view seems to entail that the natural properties that are relevant for moral and legal responsibility-attribution are related to an agent's mental capacities. That an agent's mental capacities are, *ceteris paribus,* what make a difference with regard to responsibility-attribution is a fairly common view – probably the most common among compatibilists. Often, *reasons-responsiveness* is the mental capacity that is picked out as sufficient for responsibility-attribution. In the next section, I will argue that given that the metaphysical doctrines of determinism and physicalism are accepted in this discussion, they restrict us regarding what can plausibly be picked out as a relevant natural difference that can account for a moral difference.[32, 33] Furthermore, I will suggest that determinism and physicalism provide two *prima facie* challenges to the view that reasons-responsiveness is sufficient for responsibility, in the basic desert sense.

3.4.3 Metaphysical constraints on the relevant difference condition

In the present discussion, the metaphysical doctrines of physicalism and determinism restrict what can be referred to as a relevant difference.[34] For example, if we accept determinism, we cannot explain a moral difference between two people by saying that one of them has an indeterministic property, such as a libertarian free will, and the other has not.

---

[32] As has already been pointed out, physicalism in a liberal form (i.e., not reductive physicalism) is embraced by nearly all free will philosophers. Determinism is a more controversial metaphysical thesis. The reasons for why it is accepted as a premise in the discussions of this book are described in chapter 1, where I also shortly discuss some different accounts of libertarian free will.

[33] For a definition of basic desert, see section 2.3.2 (pp. 35-49.)

[34] For a brief discussion of how these metaphysical doctrines are interpreted in this thesis, see sections 1.3 and 1.4 (pp. 15-20.)

In the following two sections, I will attempt to spell out how the metaphysical doctrines of physicalism and determinism seem to provide two prima facie challenges for the prospect of justifying basic desert on compatibilist grounds.

The challenges will then further be discussed in the next chapter, in the light of some compatibilist theories of free will and (or) responsibility in order to consider whether these theories can meet the prima facie challenges for the prospects of justifying basic desert retributivism on compatibilist grounds. [35]

3.4.4 The challenge from determinism

According to Morse, criminal law, and hence legal retributivism, presuppose that behavior can be explained by mental states such as beliefs, desires, intentions, volitions, and plans. Moreover, he claims that "the law's view is that people are capable of acting for reasons and are capable of minimal rationality" (2015, p. 255). What is required for its practice to be justified is that these presuppositions are also true: if it is true that people have beliefs, desires, intentions, volitions, and plans, and if it is true that people are capable of acting for reasons and are capable of minimal rationality, then legal practices are justified.

The capacities to be able to act for reasons and to be minimally rational seems, in Morse's view, to be necessary conditions for being responsible, and deserving of punishment. According to him, having these capacities entails than one can act freely: "it is entirely consistent with determinism that people can act freely in the sense of doing what they choose to do based on their reasons for action and doing so without compulsion" (2013, p. 39). Given that determinism is a

[35] As will become clear in this discussion, certain attempts to meet the challenge from determinism for the prospect of basic desert is via free will compatibilism, others are via responsibility compatibilism. These compatibilist accounts do not necessarily coincide. However, for the present discussion, it is not of central importance if basic desert is defended via responsibility compatibilism or via free will compatibilism. Regardless of which of these compatibilist position that is provided as the framework within which basic desert can be had, the Principle of Relevant Difference will require that there is a relevant difference with regard to the base property of basic desert and other natural properties.

metaphysical restriction that Morse accepts, what it means to act freely in the sense he has in mind has nothing to do with libertarian freedom. Rather, freely willed actions are distinguished from non-freely willed actions by the mental capacities the agent has at the time of action.

In short, the challenge from determinism is as follows: Assume that a certain mental capacity, as e.g., the capacity to act for reasons, has been suggested as the relevant difference between people who can deserve (in the sense of basic desert) to be punished and those that cannot. Assume also that two people (say, P & Q) perform similar acts, but that only one of them has this mental capacity. Given that determinism is true, P and Q were *equally determined* to act as they did, regardless of their different capacities. So how, then, can the mental capacity in question be a relevant difference between P & Q?[36]

Elaborating on the challenge from determinism

The capacity to act for reasons is central in Morse's view, although he does not describe in detail what this capacity consists in, or what reasons really are. A commonly drawn distinction in the contemporary literature on reasons is between normative and motivating reasons (see, e.g., Alvarez, 2017). T. M Scanlon (1998) characterizes normative reasons as something that "counts in favor of" someone to act in a certain way. A motivating reason is a reason for which someone does something, and thus it is a reason that has a certain explanatory power with regard to the question "why did you do that?". A common view is that normative and motivational reasons are of different kinds: normative reasons are facts, whereas motivational reasons are mental states (see e.g. Scanlon, 1998, Mele, 2003, Audi, 1993). There are several intriguing questions connected to what reasons are, their relation to action and their role in action-explanations, but for the discussions in this thesis, it suffices to think

[36] As a reminder, basic desert is not attributed because it has good consequences (see section 2.3.2). If consequences were part of why desert is attributed to people, what counts as a relevant difference would change. Then a relevant difference between two people may be that one of them is responsive to certain kinds of interventions, but the other is not, and therefore it makes sense to attribute desert to one who is sensitive to interventions but not to the other one.)

of reasons as something an agent thinks of (or can think of) as something that speaks in favor of an action, when she deliberates about how to act. The distinction between normative and motivational reasons is helpful, since both these kinds of reasons figure in, but play different roles, in the assessment of whether someone is reasons-responsive.

Agents who act freely are, in Morse's view, doing what they choose to do based on their reasons for action and doing so without compulsion or coercion. Agents who suffer from mental disorders that make them unable to respond appropriately to rational consideration cannot, according to this view, act out of free will and are hence not responsible for their actions. I take this general description to correspond to the characterization of reasons-responsiveness as it is discussed by e.g., John Martin Fischer (1994) and Fischer and Mark Ravizza (1998), whose account has played a central role in the discussions of reasons-responsiveness and moral responsibility. Fischer and Fischer & Ravizza provide a much more detailed description of reasons-responsiveness, but for our purposes this general description of what it means to be reasons-responsive will suffice.

In discussions regarding responsibility and desert, reasons-responsiveness is often assumed to play a central role for desert-attribution, since to be reasons-responsive means that one has the mental capacity to take normative reasons into account when one rationally considers what to do. For example, even though someone thinks she needs to drive really fast because she is late for a meeting (a motivational reason) she knows that she drives on a road with a certain speed-limit and considers this normative reason as a reason to not drive as fast as she wants. In contrast, someone who is psychotic and thinks she is driving on an intergalactic highway with no speed limits seems to lack the capacity to adapt her speed to the speed-limit on the actual (real) road.

Reasons-responsiveness in this sense is obviously relevant when motivating why we have legal rules in the first place. As Morse puts it, the law is an action-guiding enterprise, and it would not work as that if people are unable to use laws as reasons for their actions. If people are reasons-responsive in the sense that they, as a general

ability, can use laws as action-guiding, it makes sense to provide reasons in forms of laws to promote certain behavior and inhibit other kinds of behavior in society at large. However, the claim that reasons-responsiveness is relevant to the law in this consequentialist sense is distinct from the claim that reason-responsiveness is relevant for basic desert. And since basic desert is required for the justification of legal retributive punishment, reasons-responsiveness can not only be of consequentialist value for the legal system: it must also be a property that in some way can lead to, or be the base- property of, basic desert. Otherwise, reasons-responsiveness fails to account for a justification of retributive punishment.

According to Morse, someone can deserve punishment (or praise, blame, reward) in virtue of what kind of mental capacities she had when she acted. If she had the mental capacity of reasons-responsiveness she deserves punishment (or praise, blame, etc). If she was not reasons-responsive, she does not deserve it. When the moral property of basic desert is attributed to someone, the base-property is reasons-responsiveness. If someone commits a criminal action, in principle, she does not deserve punishment if she is not reasons-responsive, and she does deserve punishment if she is reasons-responsive.

If reasons-responsiveness is the base-property of basic desert, this entails that reasons-responsiveness is the *relevant difference* between someone who has the moral property of basic desert and someone who lacks that property: reasons-responsiveness is the natural property that makes a difference with regard to if you have the moral property of basic desert or not, when you commit a criminal action. I will now explain how determinism is a challenge for this view.

Consider a case in which a person P sticks to the speed limit for the road she is driving on since she takes it as a reason for not driving faster even though she is late for a meeting (i.e., even though she also has a motivational reason to drive much faster). The day after, P is late for another meeting. Still, she sees the speed limit as a reason not to drive too fast, but this time it does not override her motivation to drive faster than the speed limit allows for: the meeting is too important, and her boss was unhappy with her late arrival the day before. P consequently drives much too fast when she is observed by

a police officer who stops her. As it happens, another person, Q, also drives much too fast on the exact same road just behind P. Q believes she is late for a meeting with Darth Vader and believes she has good reason to hurry up since Mr. Vader is a busy person. Q is also stopped by the police officer.

P and Q are – probably – assessed differently when it comes to the question of whether they deserve legal punishment or not for what they have done. Imagine that the following line of reasoning is offered to explain this difference: P deserves punishment and Q does not, since P was reasons-responsive in the relevant sense at the time of action (t_a), which means that she is ascribed the moral property of basic desert which, in turn, makes it justified to punish her with reference to desert. Q, on the other hand, suffered from a psychosis at t_a, which means that she was not reasons-responsive in the relevant sense at t_a and since she was not reasons-responsive, she does not deserve punishment. The difference between P and Q can be spelled out as follows:

> P: Drives too fast at t_a, reasons-responsive at t_a→ deserves punishment
>
> Q: Drives too fast at t_a, ¬ reasons-responsive at t_a → ¬ deserves punishment

Reasons-responsiveness is, in this situation, supposed to be the natural property that meets the requirement of the Principle of Relevant Difference: reasons-responsiveness is supposed to account for the moral difference between P and Q. What makes it the case that P was reasons-responsive is that she, at t_a, had the ability to rationally consider the normative reason (i.e., the speed limit) and respond to it, even though she actually did not do so. Reasons-responsiveness is, hence, a dispositional property (it is defined in counterfactual terms).

In a determinist framework, all that happens do so out of necessity: given the history and the laws of nature, it could not be in any other way. In such a framework, to say that someone has a dispositional property at t_a, is to say that the person has a property such that under

other conditions than those that actually obtain at t_a (which means: had the history and/or the laws of nature been different or, alternatively put, in another possible world) then this dispositional property had been realized. In the case of P, in the actual world, she was determined to drive too fast, but if the world had been different with regard to the history and/or the laws of nature, i.e., in another possible world, then she would have adapted her speed in accordance with the speed limit.

The challenge from determinism is that given that determinism is true, P and Q were *equally determined* to produce the same functional output at t_a, regardless of their different functional outputs under other conditions. But it is difficult to make sense of why the fact that P has a dispositional property such that she would have adapted her speed under *other* conditions, would constitute a relevant natural difference with regard to natural properties between P and Q at t_a .Why think that the dispositional nature of their mental abilities is a relevant natural difference between them – i.e., different such that P deserves to be punished and Q does not – when it is true of both P and Q that they were completely determined to act as they did (and to have their respective motivational reasons) in the actual situation?[37]

---

[37] Again, I want to remind the reader that it is important to keep in mind that I am discussing *basic* desert. We can talk of desert in other senses. For example, we may attribute desert because it has good consequences to do so. But then it is not *basic* desert we are talking about. As will be discussed to some extent in chapter 7, that kind of desert is not be based on intrinsic properties of an agent, in the same sense as basic desert, or responsibility in the basic desert sense, is. On a consequentialist view of desert, the difference between a person who deserves punishment and one who does not is a difference regarding the consequences of punishing them. Whether a person deserves to be punished depends on some non-intrinsic property of her, namely whether she is the kind of person that it has good consequences to treat in certain ways (or something similar). One might argue that there is such a difference between e.g., people like P and people like Q – and if there is, then this difference is there even if they are both determined to act as they do. But since my discussion concerns basic desert specifically, what counts as a relevant difference between people is different – there has to be an intrinsic difference between the two persons such that it becomes plausible to say that one is a person who can deserve – in the basic desert sense – and the other is not. The challenge from determinism then claims that the fact that two persons were equally determined to perform their acts makes them sufficiently similar, why it would violate the consistency requirement to attribute basic desert to one of them but not to the other.

The challenge from determinism is not intended to prove that reasons-responsiveness does not work as the relevant natural difference with regard to basic desert. Rather, it is a way to describe a common prima facie worry whether compatibilism can deliver an intuitively plausible account of basic desert and hence justify legal retributivism. I will discuss this worry in relation to specific compatibilist arguments in next chapter. Before that, I will describe a second challenge provided by our metaphysical commitments: the challenge from physicalism.

3.4.5 The challenge from physicalism

The second challenge for the prospect of justifying legal retributivism on the basis on free will compatibilism comes from the metaphysical constraints of physicalism, and more specifically, physicalism of the mental.

It is uncontroversial that the difference between people who deserve to be punished for actions they have performed (in the basic desert sense) and those that do not, has to do with how their actions came about, e.g., whether they could control their acting in some relevant sense. Compatibilists who cannot refer to libertarian free will as the relevant difference in this regard must instead say that the relevant difference has to do with some mental capacity of the persons, a mental capacity involved in the production of the actions in question, such as reasons-responsibility.

The challenge from physicalism builds on the following plausible extension of the Principle of Relevant Difference. Suppose that we are considering the suggestion that reasons-responsiveness is the relevant difference between someone who has basic desert and someone who does not: it is, as it were, the desert-base. We can then further ask about other mental capacities that we think do not plausibly qualify as desert-bases: what is the relevant difference between reasons-responsiveness and these other capacities? If we cannot find a relevant difference, a difference that makes it plausible to think of one but not the other as a desert-base, then this is a problem for the suggestion that reasons-responsiveness (and not the other capacities) is the desert-base.

In more general terms, it seems plausible to maintain that if a certain natural property G prima facie seems to be the base property for a moral property F, while another natural property G' prima facie does not appear as a plausible base property, then there must be a relevant difference between G and G' in order to make it intuitively plausible that G is a supervenience-base for F and G' is not. If no such relevant difference is to be found, it seems as if the consistency requirement is violated.

The challenge from physicalism is, in short, that given physicalism, there are no such relevant differences between, on the one hand, e.g., reasons-responsiveness (or other compatibilist suggestions), and, on the other hand, other mental capacities that, plausibly, are not desert-bases.

Let us assume that reasons-responsiveness is the base-property of basic desert (e.g., in a case where two persons perform similar actions and one of them deserves to be punished for the action and the other does not, reasons-responsiveness is the relevant difference between them.) Furthermore, assume that color-responsiveness is not relevant for basic desert. In order for these two capacities to differ with regard to basic desert, there must be a relevant difference between reasons-responsiveness and color-responsiveness that makes it plausible that reasons-responsiveness, but not color-responsiveness, is the base-property of basic desert. It might seem obvious that there are such differences. To start with, reasons-responsiveness makes a difference with regard to how actions come about in a way color-responsiveness does not. More specifically, it might be suggested that reasons-responsiveness makes a difference with regard to how actions come about compared to how actions come about when someone is *not* reasons-responsive in the sense that a person can deliberate rationally about, and thereby *choose* how to act, instead of acting upon, e.g., irresistible impulses, delusional beliefs or something like that.

Being a physicalist, one is not committed to the thesis that the capacity to deliberate rationally about reasons for action is reducible to brain functions. But one is most plausibly committed to the thesis that if two people differ in their capacity to deliberate rationally on reasons for actions, this difference must be reflected by, and depend on, differences in their brains. The point can be illustrated by an

example from legal practice. When assessing legal responsibility, offenders are at times screened for psychiatric disorders. Deviance in the brain structure may indicate that the offender lacks the ability to form her decisions for action in the way that is required for basic desert. Consequently, it seems that a person's brain must have *certain physical structures* intact in order for reasons-responsiveness to be in place. This means further that certain brain structures are at least a necessary condition for the ability to be reasons-responsive (for the moment, I leave the discussion of multiple realizability aside).

Given physicalism, then, the difference between a person who acts out of reasons-responsiveness and one who does not, is a difference between which brain-structures (or brain-processes) that were involved in producing their actions. The suggestion that reasons-responsiveness (but not other capacities) is the basis for desert attribution, in effect entails that the involvement of certain brain structures (or brain processes) in producing action, but not the involvement of others, is the basis of desert.

However, if we look at the level of brains, what could make it the case that a certain brain structure can be relevant for basic desert, whereas other brain structures are not? If we just look at these brain structures – regardless of their respective functions – it seems odd to say that a certain brain structure, but not another, is a base-property for a specific moral property just in virtue of its physical properties. Brain structures are just complex networks of neurons, that are wired in certain ways, and fire in certain ways. This goes for all structures in the brain. Thus, with regard to physical properties, it seems difficult to pick out a relevant difference that can account for such a moral difference.

If we compare the physical structures in the brain that are involved in a deliberative, rational decision-making (i.e., the kind of decision-making that reasons-responsive people are capable of) with the physical structures in the brain that are involved in the process that leads to action when reasons-responsiveness is lacking, it is difficult to see the relevant differences between these structures, at least with regard to basic desert. In other words: if reasons-responsiveness is the base-property of basic desert, there must be a relevant difference between reasons-responsiveness and other mental capacities that are

not base-properties of basic desert. But if we are looking at the brain structures that are involved in reasons-responsiveness, these structures are not relevantly different compared to other brain structures.

Now, it might be argued that it is not brain structures as such, but rather brain processes, or brain events, that are relevant for reasons-responsiveness, and hence for desert-attribution. However, the same reasoning that holds for brain structures holds for processes in the brain. If the physical events that lead to action in a reasons-responsive person are sufficiently similar to physical events that lead to action in someone who is not reasons-responsive, then, arguably, the consistency requirement implies that these processes should not differ with regard to moral properties (again, with regard to basic desert, so this means: either both, or none, have basic desert).

The idea is, then, that it is hard to get a sense of a relevant difference for whether someone deserves or does not deserve to be punished for her actions, if we look just at the level of brain structures involved in her acting. But given physicalism, this level is the one we should focus on, since the relevant differences are differences in how different acts come about, and it is at the physical level of brain structures, processes, and events that our actions come about.

Elaborating the challenge from physicalism

In light of the above reasoning, we can ask whether it is still intuitively plausible that reasons-responsiveness is sufficient as base-property for basic desert given that reasons-responsiveness (which in this section is assumed to be the base-property of basic desert) is sufficiently similar to other mental capacities in terms of how they are realized on the level of brain processes. In this section, I will suggest that it is not. Instead, I will argue that the fact that brain processes that lead to action in someone who is reasons-responsive are sufficiently similar to brain processes that lead to action in someone who is not reasons-responsive provides a prima facie reason to doubt that reasons-responsiveness is sufficient as base-property of basic desert.

When two people perform similar actions, but one of them fulfils the criteria for basic desert and the other does not, this moral

difference is due to a difference in the nature of the mental states (i.e., the natural properties) that are causally involved in the chain of events leading up to the respective action. If reasons-responsiveness is picked out as the relevant difference between someone who has basic desert and someone who has not, then reasons-responsiveness must, given the Principle of Relevant Difference, differ in a relevant sense from natural properties that are not base-properties of basic desert. But what is the intuitive difference between reasons-responsiveness and another property, let us say, for simplicity, the property of being non-reasons-responsive? A plausible answer is that the relevant kind of reasons-responsiveness is a capacity that consists in a responsiveness to normative reasons, and if one has that capacity, one can consider normative reasons when one chooses how to act. In other words, reasons-responsiveness makes a difference when one elaborates on different alternatives of action. If so, the relevant difference between someone who is reasons-responsive and someone who is not is described in terms of the functional role of reasons-responsiveness.

In order to provide for a functional difference, reasons-responsiveness must be a property that can be involved in mental events that cause actions. In other words, reasons-responsiveness must be causally relevant for actions.[38] How to understand notions like "cause," "causal relevance," and "causal effect" in relation to philosophy of mind and philosophy of action is much discussed. If we embrace reductive physicalism, mental states are reducible to brain states, and hence the causal role played by mental states is identical with the causal role played by brain states. However, the fact that mental states seem to be multiply realizable is a reason to reject reductive physicalism: e.g., the mental state of fear is possibly shared also by creatures with different brains compared to the human brain, which means that "fear" cannot be reduced (at least not on type-level) to a certain brain state. The causal story of mental states becomes even more complex when mental states are regarded as non- reducible

[38] Not everyone agrees that mental causation is necessary in order for mental states to be relevant for action explanations. For example, Helen Beebee (2017) argues that epiphenomenalism about mental states must not obviously a crazy position. I return to her argument for why this is so in chapter 5.

to physical states. One central worry is to what extent the so-called "Exclusion Problem" is a threat for non-reductive physicalists (see, e.g., Beebee, Hitchcock & Price, 2017). This problem can be outlined as follows: if we take mental properties to be irreducible to physical properties, and actions can be fully explained in *both* mental and physical terms, it seems as that we have two complete, but different, explanations of the same phenomenon. The problem is that these two different explanations both are regarded as *causally sufficient* for accounting for the action in question. According to Kim, when we are faced with two causal explanations of a single event, we have some different alternatives of how to think of it:

> […] (a) each is a sufficient cause and the effect is causally overdetermined, (b) they are each necessary and jointly help make up a sufficient cause (that is, each is only a "partial cause"), (c) one is part of the other, (d) the causes are in fact one and the same but given under different descriptions, (e) one (presumably the mental cause in the present case) is in some appropriate sense reducible to the other, and (f) one (again the mental cause) is a derivative cause with its causal status dependent in some sense on the neural causes, *N*. (Kim, 2000, p. 65)

In either case, Kim argues that "the presence of two causal stories, each claiming to offer a full causal account of a given event, creates an unstable situation requiring us to find an account of how the two purported causes are related to each other. This is the problem of 'causal/explanatory exclusion'" (Kim, 2000, p. 65).

There are scholars who reject the claim that there is an exclusion problem, and I will discuss an objection to the exclusion problem put forward by List & Menzies (2017) in chapter 5 (see also e.g., Burge, 1993, and Yablo, 1992, for critical discussions). However, in the main part of this thesis, I will proceed from the assumption that the exclusion problem is a challenge for non-reductive physicalism.[39]

[39] When discussing the exclusion problem, I have the metaphysical version of this problem in mind, that is, that it is problematic if effects can have two independent causes that both are sufficient causes of the effect. However, there are those who challenge this view, arguing that we should reverse the order of investigating causal relations: instead of asking wether we capture the *real* causal relations in our causal explanations, we should start with clear cases of causal explanations in order to

Moreover, I will assume, following Kim, that when we are faced with two causal explanations of a single event, alternatives (a), (b) and (c) are "nonstarters" (Kim, 2000, p. 65). Regarding the other alternatives, (d) – (f), all of them locates the causal relevance of reasons-responsiveness on the physical level. More specifically, I will assume that the relevant physical events, in this case, are physical events in the brain.

When compatibilists point to reasons-responsiveness as sufficient for moral responsibility, the intuition underlying this point is, arguably, that if a person can respond to the relevant reasons in a given situation, she has the ability to control her behavior. In contrast to someone who is not reasons-responsive, she is able to act appropriately based on the relevant reasons. If she, despite this ability, does not act appropriately, (i.e., does not respond to the appropriate reasons despite that she has the ability to do so) then it seems that she is responsible for her actions in a way that a non-reasons-responsive person is not, and thus there is relevant difference between them with regard to basic desert. The difference is that there is some element of control (i.e., the ability to choose based on reasons) involved when one of them acts, but not when the other acts. But when we look closer at this difference in light of physicalism, what previously appeared to be a relevant difference seems to disappear. Because in both cases, brain processes (and input to the brain processes) cause the actions – and neither of them involves an element of control in a relevant sense since both are just neural input-output mechanisms. Hence, when we zoom in on the physical processes involved in the actions, these processes are, in relevant aspects, not different from each other, and the relevant difference between the agents and their actions is lost.

This argument does, of course, not preclude the possibility that neuroscientists may discover a difference between brain processes that lead to action when someone is reasons-responsive, on the one hand, and processes that lead to action in someone who lacks reasons-

---

understand how the concept of causality works (see e.g., (Burge, 1993; Wilson, 1999). I will briefly discuss an objection to Kim's characatization of the exclusion argument put forward by List & Menzies (2017) in chapter five.

responsiveness, on the other hand, that will appear as an intuitively relevant difference. However, it is hard to imagine what such difference would consist in, and it is still the case that the burden of proof lies on the compatibilist who defends basic desert. Moreover, current legal practices cannot be justified with reference to what we may come to know about brains and minds in the future, but have to be justified with reference to what we currently know.

It is important to point out two things about this prima facie challenge: Firstly, this challenge is not meant to presuppose reductive physicalism. In only assumes that as far as we accept physicalism, physical effects have physical causes. I will return to this issue in chapter 5, where I will discuss a suggestion to the effect that compatibilism about desert can be saved if we accept a non-reductive physicalism. Secondly, this challenge is not a challenge for all kinds of moral property attribution. What intuitively is a relevant difference vary between different moral properties. The reason that physicalism about the mental raises a challenge for basic desert is that what might seem like intuitively relevant differences with regard to desert is found in the mental realm – it concerns how different acts are caused, and more precisely that some sort of control is involved in the causation of some acts but not others – and seems to disappear when we look at fundamental physical nature or basis of the mental processes involved. It is not at all obvious that physicalism threatens what intuitively appears to be the relevant differences when it comes to other moral properties, such as goodness, kindness, generosity, rightness and wrongness.

One of the central concerns in the vast literature concerned with moral responsibility, regardless of whether we take reasons-responsiveness, or some kind of practical rationality, or something else, to be the base-property of basic desert, is that these capacities are dispositional in nature (as all capacities are). The nature of capacities is often discussed in modal terms: it is described as something the agent having the capacity "could have done," or "can do," given certain circumstances. But in a deterministic framework, this cannot be interpreted literally: given determinism, no one can act otherwise than she actually does, and no one can have another brain than she actually has. However, there are several intriguing

compatibilist responses to this worry. I will discuss some well-known approaches in the next chapter, and consider whether they can meet the challenges from determinism and physicalism.

3.5 Summary & conclusions

In this chapter, I have described Stephen Morse's objection to the Revision Argument. Morse argues that the Revision Argument is mistaken about the role of libertarian free will in the legal system. Morse claims that legal responsibility has nothing to do with free will, but concerns the rational capacities in a person. To begin with, many people disagree with Morse that free will is not an issue in the legal context: for example, Meynen (2016) and Harris (2012) point out that there is evidence to the contrary: free will, and perhaps even libertarian free will, is an implicit assumption in legal doctrine. Moore (2016) argues that even if Morse is right about that the law from an internal aspect does not require free will, we cannot avoid addressing external questions, such as the question of determinism, free will and responsibility, in order for the internal practice of the law to make sense. Concerning such external questions, Morse claims that compatibilism can do the job of providing such an external basis to the internal practices of the law. Moreover, he claims that the law is a folk psychological enterprise, and the folk psychological account of free will and responsibility does not require libertarian free will. Thus, the Revision Argument seems to be mistaken on two points, according to Morse: (1) that the legal system presuppose libertarian free will for being legally responsible, and (2) that folk psychology is incompatibilist with regard to free will, responsibility and retributive punishment.

However, when looking into some experimental philosophy cases concerning this issue, it turned out that it is not clear whether the folk psychological view of free will and responsibility actually is compatibilist or incompatibilist. Therefore, I turned to Morse's claim that compatibilism is consistent with, and provides a secure metaphysical basis to, criminal responsibility and legal retributive punishment. I argued that in order for a compatibilist theory to account for basic desert, it must be able to meet the demands of a

moral principle that I call "the Principle of Relevant Difference." According to this principle, in order to establish the moral property of basic desert, there must be a relevant natural difference (leaning here on the ethical supervenience thesis) between someone who is ascribed with basic desert and someone who is not. Furthermore, I argued that since physicalism and determinism are accepted in this discussion, they, in combination with the Principle of Relevant Difference, constitute two prima facie challenges for compatibilist basic desert retributivism: "the challenge from determinism" and "the challenge from physicalism." I argued further that it seems difficult to pick out the relevant difference between someone who is ascribed with basic desert and someone who is not within a compatibilist framework of free will and responsibility, in light of these challenges. However, in order to assess whether they *de facto* provide problems for compatibilism with regard to basic desert, we must scrutinize some compatibilist theories in more detail.

In the next chapter I will discuss a selection of compatibilist ideas that are frequently discussed in relation to moral responsibility, and scrutinize whether these ideas can meet the challenges from determinism and physicalism, given the Principle of Relevant Difference.

4 Compatibilism, basic desert & the Principle of Relevant Difference

4.1 Introduction

In the previous chapter, I introduced the Principle of Relevant Difference according to which a claim to the effect that there is a moral difference that does not supervene on a natural difference of a relevant kind violates what Kim calls "the consistency requirement," which is grounded in the moral intuition that like cases should be treated alike (Kim, 1984, pp. 161-162). I also introduced the challenges from determinism and physicalism, which, I argue, provide metaphysical constraints of what can be referred to as a relevant difference.

According to Morse, free will compatibilism provides us with the kind of freedom required for moral and legal responsibility, basic desert attribution and, hence, also for the justification of retributive punishment. In this chapter, I will scrutinize this claim.

Morse does not specify in any detail which compatibilist account he has in mind but writes that it has to do with mental capacities rather than libertarian free will. In this chapter, I will examine some influential compatibilist theories and their suggestions of what mental capacities they claim have to be in place for someone to be morally responsible for her actions.

As already pointed out, free will compatibilism and moral responsibility compatibilism do not necessarily coincide: one can argue for free will compatibilism without being a compatibilist about moral responsibility, and one can be a compatibilist about moral responsibility while rejecting free will compatibilism as a plausible theory. Or, as will be illustrated by the parts of this chapter concerned with the writings of Harry Frankfurt, one can argue for both free will compatibilism and moral responsibility compatibilism, without arguing that free will compatibilism is sufficient for compatibilism about moral responsibility. However, besides the parts concerned

with Frankfurt's writings, I will not attend to this distinction in the following discussion. Even though free will plays a central role in the Revision Argument as well as in Morse's objection to it, whether there are any plausible compatibilist theories of free will as such is not of vital importance here. Rather, what is at stake is whether legal retributivism is justified or not and for it to be justified, we need to be able to ascribe basic desert to the people we want to punish on retributive grounds. For the sake of justifying basic desert attribution via compatibilism, we may either look for a compatibilist account of free will that gives us the kind of moral responsibility that is required for basic desert or we can try to find a compatibilist account of moral responsibility that does not necessarily address the free will question, but provides for the justification of basic desert attribution by itself. It may be worth noticing that it is not always clear to what extent different compatibilist theories actually aim to justify basic desert attribution, but there are at least not any explicit claims to the contrary: They aim to defend non-basic desert moral responsibility.

As was pointed out in section 2.5, the kind of moral responsibility in focus in this thesis is labelled, following Caruso & Morris (2017), "retributive desert moral responsibility." The notion "moral responsibility" refers to the kind of responsibility that can include, or lead to, basic desert. I will in the following discussions have this notion in mind when discussing moral responsibility, if not indicating otherwise.

If there is a version of compatibilism regarding free will and /or moral responsibility that can meet the Principle of Relevant Difference, in light of the challenges from determinism and physicalism, then we are, arguably, justified in ascribing basic desert to people and, hence, also justified in punishing them on retributive grounds. Then, Morse's claim that compatibilism can provide a secure basis of retributive punishment is accurate, and the Revision Argument can be rejected. However, if the compatibilist theories we will be concerned with in this chapter cannot meet the Principle of Relevant Difference, then Morse's claim is not supported (at least not by the compatibilist theories discussed here), and the Revision Argument stands (at least for now).

This chapter will unfold as follows. In section 4.2, the discussion starts with Harry Frankfurt's (1971) hierarchical model of free will, together with some critique of this view put forth by Gary Watson (1987). Section 4.3 is concerned with discussions about the principle of alternate possibilities (PAP). PAP is the principle that moral responsibility requires the possibility of doing otherwise than one actually did, and the question of whether this is a necessary requirement or not for moral responsibility is central to the moral responsibility debate. The section starts with Peter van Inwagen's (1983) "Consequence Argument" according to which we cannot be morally responsible for our actions if they are the consequences of factors beyond our control. Van Inwagen's argument is followed by another central contribution to the responsibility debate made by Frankfurt: his widely discussed argument that people can be morally responsible also when PAP is not satisfied. This line of thought has been developed more recently by e.g., John Martin Fischer, and compatibilist theories of this kind will be called "actual sequence compatibilism."

In section 4.4 I will discuss theories that I will call "counterfactual theories of alternative possibilities" according to which PAP does not require any metaphysical freedom, but can be satisfied if there is a counterfactual situation (i.e., a nearby possible world) in which the agent would have done otherwise. In section 4.5, the discussion moves to an alternative way of approaching the moral responsibility: Peter Strawson (1962) argues that responsibility and free will should be understood in terms of "reactive attitudes" and that the analytic disagreement between compatibilists and incompatibilists depends on a misconstruction of the nature of moral responsibility.

In section 4.6, I summarize the conclusions drawn from the discussions in this chapter: that the compatibilist theories considered turn out to have problems with meeting the requirements from the Principle of Relevant Difference, in light of the challenges from determinism and physicalism. If this conclusion is correct, it means that Morse's compatibilist objection fails to make a convincing case against the Revision Argument. But I will also introduce a possible problem for this conclusion, which will be discussed in chapter 5.

4.2 The hierarchical view of free will

The question of whether people have free will, and if so, what kind of free will they have, is most often discussed in relation to topics such as agency, desert, and responsibility (see e.g., Timpe 2010, McKenna & Coates 2015). In these discussions, human action plays a central role. Richard Taylor once wrote that "the question whether men have 'free will' is really only the question whether men ever act freely. No special concept of the will is in any way needed to understand that question" (Taylor, 1960, p. 264).

However, one reason not to restrict the discussion of free will to the possibility of free action is that our prospects of carrying out free actions are often restricted by environmental factors in the world, in a way free will in itself is not. For example, a prisoner may perhaps *want* to walk out of her prison, but she is not free to *act* in that way. Hence, we can be free to want something without having the possibility to carry out this will.

In situations where there are no particular constraints on actions, a free action can be viewed as exercising one's free will, and in such cases, the will and the action are closely connected. In the following sections, I will consider the question of whether actions that are carried out as consequences of someone's free will is *relevantly different* compared to actions that are not the consequences of someone's free will, in a way that matters to basic desert. Hence, the distinction between free will and free action will be present in the following discussion. However, in relation to the overall question – if it is reasonable to ascribe basic desert to people on the grounds that they acted freely, it is not necessary to view free will and free action as distinct categories. In order to make sense of the question about why people deserve punishment in the sense of basic desert, free will must be discussed in combination with actions that are carried out as a consequence of the free will, since basic desert is a (moral) property that people can have because of something they have done.

In his article "Freedom of the Will and the Concept of a Person" (1971), Harry Frankfurt presents his hierarchical model of how to understand free will and free action.

According to Frankfurt, the essential difference between human free will and the will in other creatures is to be found in their psychological structure. More specifically, Frankfurt thinks that human beings, in contrast to members of other species, are able to form what he calls "desires of the second order:"

> Human beings are not alone in having desires and motives, or in making choices. They share these with the members of certain other species, some of whom even appear to engage in deliberation and to make decisions based upon prior thought. It seems to be peculiarly characteristic of humans, however, that they are able to form what I shall call "second-order desires" or "desires of the second order." Besides wanting and choosing and being moved *to do* this or that, men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are. (Frankfurt, 1971, pp. 6-7)

Frankfurt suggests that a central difference between human and animal behavior is the human capacity of reflecting on her beliefs and desires and form *second-order* beliefs, desires and judgements with her own will as the object. According to Frankfurt's account, we act freely when the desire upon which we act is one that we also approve of, and want to act upon when reflecting over it.

Frankfurt's view accommodates the intuition that there is an important difference between acting out of compulsion, i.e. not being able to resist acting in accordance with one's desires, preferences etcetera, and acting in accordance with them because one chooses to do so. Frankfurt illustrates his point by taking an unwilling addict as an example: this person hates his addiction and tries everything that he thinks might enable him to overcome his desires for the drug, but always fails. In such a case, there is a first-order desire to take the drug, but there is also a second-order desire not to have the desire to take the drug. That is, the addict wishes that he was not a person who wanted to take the drug. Even if this unwelcome desire to take the drug is impossible to resist, the unwilling addict who truly identifies with his second-order volition (that has his first order desires as object) may state– he does not identify himself with the desire that he acts upon: this desire is not really *his* and wish he could resist it. If this is the case, the unwilling addict does not act freely, even if he acts

upon a desire that belongs to him, in a sense. In contrast, an addict who acts out of free will has a second-order desire that approves of her first-order desire to take the drug. She takes the drugs since she has a desire for it, she chooses to act on this desire, and make no effort no to do so.

Frankfurt's hierarchical model is intended to describe and explain what must be in place in a person's psychology for her to act out of free will. Even though all attitudes, desires, beliefs, and so on, strictly speaking, belong to the individual having them, only some of them can be the basis of freely willed actions. According to this model, if someone acts in accordance with her second order desires, the act is grounded in her as a person and her will is the source of the action, in contrast to when an agent acts upon a desire that she does not endorse as a desire she also wants to have.[40]

Determinism is not a threat to this view – Frankfurt explicitly says that freedom of will may be causally determined:

> It seems conceivable that it should be causally determined that a person is free to want what he wants to want. If this is conceivable, then it might be causally determined that a person enjoys free will. There is no more than an innocuous appearance of a paradox in the proposition that it is determined, ineluctably and by forces beyond their control, that certain people have free wills and that others do not. (Frankfurt 1971, p. 20)

Frankfurt's compatibilist theory of free will is intuitively plausible cases, such as, for example, we want to distinguish between people

---

[40] Even if Frankfurt does not mention it explicitly, his ideas relate to another common idea in free will debate, the one about sourcehood. A common intuition is that a freely willed action must have its source within the agent (see e.g. McKenna & Coates, 2015). Roderick Chisholm once defended the idea that human freedom entails an absence of causal determination, and a free agent has "a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved" (Chisholm, 2015, p. 352). The idea that someone must be 'the prime mover unmoved' is found in what is sometimes referred to as 'the Source Incompatibilist Argument,' according to which someone only acts out of her free will if she is the *ultimate* source of her action. But since no person is the ultimate source if determinism is true, no person has free will, or so the argument goes. (see McKenna & Coates, 2015).

who act upon desires that they fully embrace, and people who try to resist their desires but fail, as in the case with the unwilling addict. However, Frankfurt has received critique that suggests that his account of free will fails to satisfy the Principle of Relevant Difference which, I argue, is a requirement for being able to account for basic desert. Gary Watson (1987) puts forth an objection to Frankfurt's model that I take to be central for our purposes here. Watson criticizes Frankfurt's free will compatibilism for being arbitrary with regard to which level we should choose as the one corresponding to free will:

> The problem […] is not that there is a regressive ascent up the hierarchy, or that people are not that complex, but simply that higher-order volitions are just, after all, desires, and nothing about their level gives them any special authority […]. (Watson, 1987, p. 149)

Watson points to the fact that the capacity to form second-order desires does not imply that there is anything special about these higher order desires compared to the first-order desires, with regard to their nature. The uniqueness of higher-order desires compared to first-order desires is the content, or the object of the desire: whereas first order desires most often are directed towards the world, second-order desires are directed towards the subject's first order desires. The fundamental nature of second-order desires is, however, similar to the nature of first order desires.

Watson's critique implies that Frankfurt's free will compatibilism fails to satisfy the Principle of Relevant Difference, since there seems to be no relevant difference in nature between the different levels of volitions that Frankfurt points out as central in order to understand what it means to be free, or, as Watson puts it, there is nothing about higher-order volitions that gives them "any special authority." This point means that even though Frankfurt's compatibilism is useful when trying to distinguish different psychological mechanisms from each other, it fails to provide a secure basis for moral responsibility and basic desert. To be fair, Frankfurt did not intend his free will compatibilism to be the basis of responsibility. Moral responsibility does not, he claims, require free will:

> The most common recent approach to the problem of understanding the freedom of will has been, indeed, to inquire what is entailed by the assumption that someone is morally responsible for what she has done. In my view, however, the relation between moral responsibility and the freedom of the will has been widely misunderstood. It is not true that a person is morally responsible for what he has done only if his will was free when he did it. He may be morally responsible for having done it even though his will was not free at all. (1971, p. 18)

However, Frankfurt has made a significant contribution also to the compatibilist camp of moral responsibility, in which he attacks one of the most fundamental assumptions in the moral responsibility debate: the assumption that in order to be morally responsible for one's actions and the consequences thereof, one must have had the possibility to do otherwise than one actually did. The discussion surrounding the idea that alternative possibilities are required for moral responsibility is the topic of next section.

4.3 The Principle of Alternate Possibilities

> Why is everyone so sure they could have acted otherwise? After all, nobody ever has. (R. Taylor, cited in Yablo, 2008, p. 154)

Peter van Inwagen once argued that a necessary condition to be genuinely responsible for one's actions is that one was able to refrain from doing what one did:

> Almost all philosophers agree that a necessary condition for holding an agent responsible for an act is believing that the agent *could have* refrained from performing that act. (van Inwagen, 1975, p. 188)

This condition is often referred to as "the principle of alternate possibilities," or "PAP", in the literature. If PAP is accepted, determinism seems to pose a real threat against genuine moral responsibility. Peter van Inwagen's most discussed contribution to this line of thought is known as "the Consequence Argument," which goes as follows:

> 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 it is 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, 1983, p. 16)

Van Inwagen's claim that the consequences of our acts are not "up to us" is meant to point to the fact that we lack the necessary condition for being morally responsible for our actions, since everything that happens is the consequences of the laws of nature and events in the remote past and, hence, we have no possibility of doing otherwise than we actually do.

4.3.1 Rejecting PAP

But why think that moral responsibility depends on the possibility of doing otherwise in the first place? Prior to van Inwagen's construction of the Consequence Argument, Frankfurt put forward an objection to the idea that PAP is necessary for moral responsibility. Frankfurt presented a thought experiment in which we intuitively (or so he argues) want to hold a person morally responsible even if she could not have done otherwise at the time of action. A version of this thought experiment goes as follows:

> Suppose that someone – Black, let us say – wants Jones to perform a certain action. Black is prepared to go considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones initial preferences and inclinations, Black will have his way. (Frankfurt 1969, p. 835)

What Frankfurt aims to show with this thought experiment is that Jones is responsible for his actions *even if* the PAP is violated. This move is made by postulating that Black actually does not have to do anything because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, Frankfurt argues, "it seems clear Jones will bear precisely the same moral responsibility for what he does as he would have had if Black had not been ready to take steps to ensure that he does it. It

would be quite unreasonable to excuse Jones for his action, or to withhold the praise to which it would normally entitle him, on the basis of the fact that he could not have done otherwise" (Frankfurt, 1969, p. 836).

Many contemporary philosophers endorse Frankfurt's conclusion that moral responsibility does not require the possibility to do otherwise, and there are many more or less technical proposals of how to get around van Inwagen's consequence argument by challenging the premises and/or the conclusion. John Martin Fischer thinks that the irrelevance of alternative possibilities is intuitively very plausible, even though difficult to prove:

> One is supposed to see the irrelevance of alternative possibilities simply by reflecting on examples. I do not know how to *prove* the irrelevance thesis, but I find it extremely plausible intuitively. When Louis Armstrong was asked for the definition of jazz, he allegedly said: 'If you have to ask, you ain't never gonna know'. I am inclined to say the same thing here: if you have to ask *how* the Frankfurt-type cases show the irrelevance of alternative possibilities to moral responsibility, 'you ain't never gonna know.' (Fischer, 2002, p. 292)

Frankfurt and Fischer refute the claim that moral responsibility requires alternative possibilities: they argue that our intuitions accommodate moral responsibility attribution also without the possibility to do otherwise. They argue that what we care about in responsibility-attribution is not metaphysical unavoidability of the event as such, but how the agent came to act as she did. As Fisher puts it, "moral responsibility is a matter of the history of an action (or behavior) – of how the actual sequence unfolds – rather than the genuine availability of alternative possibilities" (Fischer, 2012, p. 124). We can call such views "actual sequence compatibilism."

Arguments of the kind presented by Frankfurt and Fischer have been widely discussed. David Widerker (1995, 2000, 2006) criticizes Frankfurt's argument for being incoherent. Frankfurt claims that his argument is supposed to show that moral responsibility is compatible with determinism, but according to Widerker, the argument in itself actually presupposes indeterminism. Because if Black really knew what Jones was going to do at a point before Jones actually did it, he would have had access to the facts about what Jones was going to do

before he did it. But how could he have access to such facts? In an indeterministic world, there are no such facts before Jones actually carries out his act. Then, it makes no sense that Black had to put in a device in Jones' brain and wait until Jones made up his mind, before interfering. On the other hand, if we agree that the world is indeterminist, even if Black intervenes in this situation, it is still the case that Jones has the possibility to choose again (see e.g., Widerker 2006, p. 165). Fischer (2002) and Robert Kane (1985, 1996) have objected along similar lines.

Structural worries like the one Widerker points out have been followed by several improvements of Frankfurt's original example. Here, I shall not discuss such improved examples, but proceed from the assumption that there are Frankfurt-style examples that avoid such criticism.[41]

4.3.2 Actual sequence compatibilism & the Principle of Relevant Difference

In relation to the questions which we are examining in this chapter – whether compatibilist theories of free will and/or moral responsibility can provide what we need in order to be justified in attributing basic desert in light of the Principle of Relevant Difference and the metaphysical restrictions of physicalism and determinism – actual sequence compatibilism seems problematic. The problem lies in the fact that according to actual sequence compatibilism the attribution of basic desert is based on a person's natural properties. As Fischer writes, "moral responsibility is a matter of the history of an action (or behavior) – of how the actual sequence unfolds – rather than the genuine availability of alternative possibilities" (Fischer, 2012, p. 124). But according to the challenge from determinism, it is difficult to pick out the relevant difference between sequences that do not lead to moral responsibility and sequences that do since all sequences are equally determined to unfold as they do. If one wants to escape this challenge by arguing that the important thing with regard to moral responsibility has nothing to do with determinism but is rather about

[41] See e.g., Alfred Mele & David Robb (1998) and Widerker (2000) for a discussions and improved examples.

the "actual sequence" that is involved when actions come about, we are instead troubled by the challenge from physicalism. When looking at the physical level, all actions seem to be caused by processes that are, with regard to their basic nature, sufficiently similar: they are neural processes that work according to input-output principles, and the way they process incoming input is completely determined by the physical and functional properties of the neural networks involved.

Besides the restrictions of determinism and physicalism, which exclude certain theories of mental states from our discussion, there are no obvious clues to how to interpret the nature of mental states in the current discussion. In section 1.6, I discussed some theories of mental states that are compatible with physicalism and determinism. If one embraces the identity theory of mind, i.e., reductive physicalism, or eliminativism, mental causation is completely explained in terms of brain states. It is more difficult to explain mental causation while subscribing to either property dualism or functionalism. The literature of mental causation is rich and complex, and I will not go into details of possible solutions for non-reductive physicalist theories with regard to mental causation here (I will discuss it to some extent in chapter five). The point that I want to make here, in relation to the challenge from physicalism and the Principle of Relevant Difference, is the following: assuming that there is mental causation, we must determine which properties in a mental event that are causally relevant.[42] Being a physicalist, and following Kim (2000) regarding the exclusion problem, the causally relevant property of a mental event is either identical to, reducible to, or otherwise dependent on a neural cause. Plausibly, this neural cause is a neural event in the brain. But there seems to be no relevant difference between such neural events in the brain that cause the kind of actions for which we are morally responsible and neural events in the brain

[42] In discussions concerning mental causation, I follow Beebee (2017) and take properties to be the relata of *general* causal relations, as for example, "mental property *M* is the cause of physical property *P*." But at the *token* level, properties are not the relata of causation, but events are. The role of properties in event-causation is that certain properties of an event are causally relevant, whereas others are not (see Beebee, 2017, pp. 287-88).

that cause other kinds of actions for which we are not morally responsible – they seem to be the same *kind* of events. Hence, to ascribe basic desert to someone on the basis of how the actual sequence of action unfolds seems to violate the consistency requirement.

4.3.3 Reinterpreting PAP

In the previous section I argued that actual sequence compatibilism cannot account for the relevant difference that is required for basic desert. However, many compatibilists accept that alternative possibilities – that is, the ability to act otherwise than one actually does – is necessary for responsibility and that there are compatibilist-friendly notions of abilities that give us that possibility. That is, there are compatibilist theories of moral responsibility according to which an agent who is morally responsible has alternate possibilities available at the time of action, in virtue of having certain abilities or capacities. In this section, I will discuss whether this kind of compatibilism fares better with regard to the challenges from determinism and physicalism, in satisfying the Principle of Relevant Difference for moral responsibility.

Dana Nelkin (2011) argues that Frankfurt's refusal of PAP as it is formulated in the case with the "Frankfurt controller" rests upon an invalid counterfactual scenario. Recall that in Frankfurt's thought experiment, Black wants Jones to perform a certain action, but he will not interfere with Jones plans if he does not have to. If Jones decides to act in the way Black wants, Black will not do anything. But if Jones decides to act in any other way, Black takes effective steps to ensure that Jones decides to perform, and then performs, what Black wants him to perform. But this move seems to preclude the relevant sense of ability. Nelkin formulates it as follows:

> We saw earlier that one way of understanding the possession of an ability is that an agent has an ability to X if (i) the agent possesses the capacities, skills, talents, knowledge and so on that are necessary for X'ing, and (ii) nothing interferes with or prevents the exercise of the relevant capacities, skills, talents and so on. Since in the Frankfurt-style examples, Joe satisfies these conditions, there is a sense in which he has the ability to do otherwise. Now, of course, there is *a* sense of "ability

> to do otherwise" in which Joe lacks such an ability. Joe will push the child off the pier, even if he wavers. In this sense, having an ability to do X is precluded when it is inevitable that the agent will not do X. (2011, p. 115).

Nelkin's critique consists in pointing out that there seems to be two different senses of "ability to do otherwise" invoked in Frankfurt's original example. When we talk about a person's ability to do otherwise, we are interested in whether the person can act differently *given that the causal structures governing her choices and actions are fixed.* But in the Frankfurt-controller case, what would have happened in the counterfactual case in which Black would have intervened, is that he would have changed Jones' causal structures, and therefore the counterfactual case alters the relevant factors in the situation. Nelkin claims that this move makes the conclusion – that PAP is not necessary for moral responsibility – invalid. A similar objection is put forth by Michael Smith who argues that the relevant ability in Frankfurt's examples are masked by other factors, but abilities should, when discussed in relation to moral responsibility, be analyzed as *dispositional properties* (more specifically, in Smith's view, this dispositional property is a certain rational capacity.) In doing so, the ability to do otherwise is present in the Frankfurt-controller example (Smith, 2003, p. 25).

Both Nelkin and Smith, then, argue that to be able to act otherwise is to have a certain psychological capacity. The general idea seems to be that this captures an intuitively relevant notion of "being able to do otherwise" which is compatible with determinism: Even if I was determined by the past and the laws of nature not to do A, there was nothing about my psychological capacities that stopped me from doing it – it was, as it were, within the range of options set by my psychological capacities to do so. In this way, these compatibilist views are descendants of more simple "conditional" variants of compatibilism. In David Hume's version of compatibilism, for example, responsibility arises from freely willed actions, which are perfectly compatible with determinism:

> By liberty, then we can only mean a power of acting or not acting, according to the determinations of the will; that is, if we choose to remain at rest, we may; if we choose to move, we also may. Now this

> hypothetical liberty is universally allowed to belong to everyone who is not a prisoner in chains. (Hume, 2019, p. 617)

The idea is that to say "I had the ability to do X" (which I did not do), is just to say that if I had tried or chosen to do X, I would have done it (see also Ayer: "It is not [...] causality that freedom is to be contrasted with, but constraint" (1954, p. 278)).

Theories of this sort, then, promise to deliver a common sensical notion of "could have done." For example, while jumping to the moon is excluded by my capacities, raising my arm is not. And indeed, intuitively I could have raised my arm ten minutes ago, but I could not have jumped to the moon (even though I actually performed neither of these actions).

In this context we should also mention David Lewis, who, even though he did not explicitly suggest any specific understanding of "ability" or "could have done", can be seen as defending such commonsense ideas of ability from one kind of objection having to do with determinism:

> I have just put my hand down on my desk. That, let me claim, was a free but determined act. I was able to do otherwise, for instance raise my hand. But there is a true historical proposition *H* about the intrinsic state of the world long ago, and there is a true proposition *L* specifying the laws of nature that govern our world, such that *H* and *L* jointly determine what I did. They jointly imply the proposition that I put my hand down. They jointly contradict the proposition that I raised my hand. Yet I was free: I was able to raise my hand. The way in which I was determined not to was not the sort of way that counts as inability. (Lewis 1981, p. 113).

It seems correct, as Lewis points out, that in some intuitive sense we are able to raise our hands, even in situations where we are determined not to. And in this sense, we are not able to jump to the moon (even though we are equally determined not to do that). Perhaps the relevant sense of ability is captured by some theory of the sort mentioned above. The objection that Lewis aims to tackle is this: even if it might seem that (in some situation) we were able to raise our hands when we, in fact, did not, if we had that ability, then we must have been able to break the laws of nature (or change the past), and of course we can do neither (Lewis admits).

Lewis' claims that we must distinguish between a person's ability to, for example, raise his hand, and his ability to break the laws of nature. It may be perfectly true to say that someone was able to raise his hand (even if he did not), while it is not true that someone has the ability to break the laws of nature. "Soft determinism" – which in Lewis' vocabulary refers to the position that one sometimes freely does what one is predetermined to do and in such a case one is able to act otherwise even though the past history and the laws of nature determine that one will not act otherwise – "[is] committed to the consequence that if I had done what I was able to do – raise my hand – then some law would have been broken" (Lewis 1981, p. 114). But, according to Lewis, this commitment can be spelled out as either a weak or as a strong thesis (p. 115):

(Weak Thesis) I am able to do something such that, if I did it, a law would have been broken.

(Strong Thesis) I am able to break a law.

Lewis argues that to be able to raise one's hand even though one is predetermined to put it down should be understood in accordance with the weak thesis: that if one would have raised one's hand, a law would have been broken. It is not the case that the hand-raising would *cause* a law to be broken. Rather, if one had raised one's hand, a law would have been broken beforehand (1981, p. 116).

Whatever we think of the success of Lewis' claim, his treatment of this kind of compatibilist friendly ability to act otherwise highlights the counterfactual nature of the ability.[43] What makes it true that I was able to raise my hand (when I, in fact, did not) is that I did so in some non-actual nearby possible world (with different laws of nature). In this world, raising my arm was, after all, excluded by the past and the laws of nature in this world.

I will call theories that make use of such a counterfactual understanding of capacities in order to save PAP for "counterfactual theories of alternative possibilities."

---

[43] See Beebee (2003) for an argument to the effect that he does not succeed in distinguishing between the weak and strong thesis in a way that ensures that we are never able to act otherwise in the sense spelled out by the strong thesis.

4.3.4 Some worries concerning counterfactual theories of alternative possibilities

Before going into the details of whether counterfactual theories of alternative possibilities satisfy the Principle of Relevant Difference and hence justify basic desert attribution, I will mention van Inwagen's worry about the expressions "could have done" and "is able to". In "Some Thoughts on An Essay on Free Will" (2015), he discusses the ambiguous meaning of "could have":

> "Could have" is grammatically ambiguous, and this has caused a great deal of confusion in the discussions of the free-will problem in English[...] Sentences of the form "X could have done Y" can mean either "X might have done Y" (i.e., "This is how things might have turned out: that X did Y") or "X was able to do Y" [...] One of the confusions that has resulted from the double meaning of "could have done otherwise" is that some critics of libertarianism have supposed that when libertarians say (for example), "She was not morally responsible for what she did because she could not have done otherwise" they mean "She was not morally responsible for what she did because her act was determined to occur". Now libertarians do believe that "X was able to do Y" entails "X might not have done Y" (i.e, "The world as it was just before X did Y might have evolved in such a way that X did not do Y"), but they regard this as a substantive philosophical thesis. They do not regard "X was able to do Y" and "X might not have done Y" as two ways of saying the same thing [...] Suppose, for example, that Martha Argerich is stranded on a desert island (where, of course, there is no piano). Is she able to play Pictures at an Exhibition? In one sense of "able to play" she is (she knows it, as the idiom has it, forwards and backwards), and in another sense she is not. I would explain the relevant sense of "able" in terms of what is presupposed by making a promise: if a fellow castaway begs Argerich to play Pictures (then and there), she is not in a position to promise to do so, for (in the sense of "able" that is relevant to the problem of free will), she would not be able to keep that promise. (van Inwagen 2015, pp.17-18)[44]

[44] Van Inwagen (2015, p. 17) complains that the free will debate has, to a large extent, been captured by "verbal essentialism". The disagreement in the debate is often centered around verbal disputes about how to interpret fundamental concepts and expressions. This problem would be interesting to discuss in light of some theories of disagreement, such as, for example, Plunkett & Sundell's "Disagreement

Van Inwagen complains that if the compatibilist needs to understand "could have done" in a counterfactual fashion, this understanding does not capture what we intuitively have in mind when talking about free will and responsibility. In the Martha Argerich case, it is true that there is nothing about her psychological abilities that makes here unable to play the piano – she could have done so if she would have had a piano. But she does not have a piano, so this counterfactual truth does not mean that she has a real access to the possibility of playing the piano. Likewise, in some reasonable sense of ability, nothing about Lewis' psychological abilities stopped him from raising his arm (if the laws of nature had just been a tiny bit different, he would have used that ability to do so). But still, he was determined by the past and the laws of nature not to do so – so the counterfactual truth that he would have done so if the laws of nature had been slightly different does not seem to mean that he has a real access to a possibility of raising his arm.

However, even if one would, in the end, think that the counterfactual view captures some sense of free will, and perhaps even some sense responsibility, one can object to the idea that it is sufficient for basic desert. Nicole Vincent expresses this worry (but she puts in in terms of justification of blame rather than basic desert) as follows:

> [...] when we remind ourselves that under determinism nobody would have capacities in the genuine access to alternative possibilities sense, the fact that in these other senses people might still have capacities seems to lose its ability to provide a normative justification for blame. After all, "blame" is a contrastive notion in the sense that it compares the way that a person actually behaved to a norm which specifies how they should have behaved – i.e. one is blameworthy when one's actions do not accord with (when they infringe) how one ought to have acted. However, a consequence of the ought implies can maxim is that how a person ought to act depends at least in part on how they can act – i.e. what a person should do depends at least in part on their capacities.

---

and the Semantics of Normative and Evaluative Terms" (2013) in which they discuss different levels of disagreement in moral matters, between participants who do not share the same metalinguistic intuitions. I lack the space in this thesis to examine how such a view could be applicable to the free will debate but I think it would be an interesting approach to the free will problem in future research.

> Stated schematically, a person is blameworthy when they acted not in accordance with how they ought to have acted, and how they ought to have acted depends at least in part on how they had the capacity to act. However, the sense of "capacity" in which people must have a capacity to act in a given way in order for it to be legitimate to expect them to act in that manner – in order for it to be true that they ought to act in that way – is surely the genuine access to alternative possibilities sense. It must surely be such that someone really could have acted in a way that accorded with how it is claimed that they ought to have acted and not just that they "could" have acted in that way. If people only ever have capacities in these other senses but not in the genuine access to alternative possibilities sense – if nobody can ever do anything other than what they actually do – then how can it ever be morally legitimate to expect anyone to do something other than what they actually did, and thus to blame them when their actions depart from that expectation? A substantial sense of "capacity" seems to be needed to warrant blame, for otherwise what we will end up saying is that despite the fact that a particular behavioral option was never in fact available to someone – despite the fact that they only "could" have done it, but not that they could really have done it – they still nevertheless ought to have done it. (Vincent, 2013, pp. 183-184)

Vincent is worried about how we can attribute blame on good grounds, if we accept that nobody can ever do anything else than they actually do. Her concern is based in the intuitive plausibility of the Kantian moral principle "ought implies can", which basically is what grounds the intuition of why PAP is important for moral responsibility in the first place.

The content of Vincent's worry about the justification of blame is similar to the worry about the justification of basic desert. Blame and desert are closely connected: if someone does not *deserve* blame, it seems unjust to blame her. Vincent is also, even if she does not express it in these terms, concerned with the question of how a mere dispositional capacity of being able to act differently can be the *relevant difference* that distinguishes a person who is blameworthy from someone who is not. Now, this is exactly what was pointed out as a prima facie problem for basic desert in relation to the challenge from determinism.

As a reminder, the problem is this: given determinism, no capacity can be such that it in a specific point in time *actually* can produce another output (as e.g., another action) than it actually does. Given

the moral supervenience thesis, moral properties supervene on natural properties, and given the Principle of Relevant Difference, there must be a relevant difference between two natural properties in order for there to be a difference in moral properties. But how could a feature of a capacity, say, the *dispositional* feature of being able to produce another output, work as the relevant difference, when this feature, in fact, could not be realized under the circumstances at hand? When presented like this, the challenge from determinism and Vincent's worry appear to be concerned with the same problem: is a dispositional capacity to act otherwise sufficient for basic desert and blame? However, as Vincent also discusses in her article, her worry, that *ought* implies *can,* can be answered by those who defend the dispositional interpretation of "can" by saying that as a matter of fact, a person who has the capacity to act differently in a counterfactually sense can act differently *in the relevant sense*. Neil Levy notes that it seems perfectly true to say about e.g., a vase that it has the property of fragility, and that this property is one the vase possesses at the time of the utterance in the actual world. This means that it is true to say that a vase *can* break, even though it for the moment being sits on the shelf:

> Fragility is a *property* the vase possesses right now, while it sits on the shelf [...] Fragility is [...]an actual sequence property, which the case has *now,* regardless of what might happen in nearby worlds [...] We test for fragility [...] by asking how [the vase] would behave in a world in which the vase was dropped [...] because doing so is the best way to understand the properties the vase has in *this* world. (Levy, 2008, pp. 227-228)

According to Levy, to understand dispositional properties in terms of possible worlds is the only way to understand the notion of capacities of mechanisms in deterministic settings (Levy, 2008, p. 228). However, Vincent objects to this claim, not primarily because it does not capture what we mean by dispositional properties, but because it cannot account for the normative role dispositional properties are assumed to play in a theory of responsibility. Vincent' s worry is, basically, that a counterfactual analysis of capacities of the sort suggested by e.g., Levy cannot meet the challenge from determinism:

> "Capacity" is ultimately a modal notion, and it is tricky business to figure out how to accommodate modality under determinism. Perhaps more importantly, however we decide to make sense of that notion – i.e., whatever analysis we propose – we must ensure that in the context of debates *about responsibility* someone stuck in a deterministic universe will have no legitimate grounds for complaints of the sort I aired above [that "ought implies can" and that one lacks the relevant sense of 'can' in an analysis of the sort suggested by e.g., Levy]. (Vincent, 2013, p. 188)

In order to solve the problem of how a notion of capacity can provide us with the normative force needed for responsibility and blame, Vincent provides an alternative, *diachronic* account of how to understand what it means to have a capacity, in contrast to e.g., Fischer & Ravizza's *synchronic* account, (see Vincent, 2013, pp. 188-191). In Vincent's view, we should understand the modal nature of a capacity, as e.g., one's capacity of being reasons-responsive, in terms of how one's reasons-responsive mechanism behaves *"in this world over a span of time"* (Vincent, 2013, p.188). According to Vincent, this approach has several advantages, but for the current purposes, what is interesting is how such an approach can meet the challenge of determinism. It turns out that it does so by abandoning basic desert, and instead justifies the responsibility practice that connects to judgments of capacities with reference to consequentialist concerns: " […] holding some people responsible for what they do but not others might be morally justified […] because doing so will be fair as long as over the course of a lifetime that person stands to benefit more than they stand to be burdened from such a practice" (Vincent, 2013, p. 189). However, the purpose of this chapter is to scrutinize whether there is a compatibilist view of free will and /or responsibility that can save basic desert – and Vincent's account does not save that.

In my view, Vincent has an intuitively plausible point regarding the tricky business of making sense of capacities in a deterministic world. In the next section, I will scrutinize the counterfactual account of alternative possibilities, and see how it fares in light of the Principle of Relevant Difference and the challenges from determinism and physicalism.

4.3.5 Counterfactual theories & the Principle of Relevant Difference

The Principle of Relevant Difference poses the following question to counterfactual theories of alternative possibilities: what is the *relevant difference* between a person who has the property of reasons-responsiveness and a person who lacks this property, with regard to moral responsibility (in the sense of retributive desert moral responsibility)? As we have seen, the properties that account for the relevant difference between someone who is and someone who is not morally responsible are, in counterfactual theories of alternative possibilities, found in the *dispositional properties* of an agent. Such a dispositional property is, for example, that at a specific time t1, when you choose coffee instead of tea, it is true to say that the mechanism that was causally involved in the choice at t1, also had the disposition in t1 to be causally involved in another choice, *had the circumstances been slightly different*.

Reasons-responsiveness is a dispositional property that is often discussed as being sufficient for basic desert. Imagine two people, A and B. A is reasons-responsive, whereas B is not. This difference can be illustrated as follows:

	t1'	*t1**	*t1@*
A:	x	y	z
B:	z	z	z

t1'- t1@ is the same point in time, in slightly different possible worlds. We approve of A's x'ing and y'ing in t1' and t1*. We do not, however, approve of her z'ing in t1@. B is z'ing in all cases (we disapprove of B's behavior in all cases.) Further, A's action in t1' and t1* are due to her responding properly to reasons in the situations, and therefore she is described as reasons-responsive. Since B does not respond properly to reasons in any of the possible worlds, she is described as not reasons-responsive.

t1@ is t1 in the actual world, t1' and t1* are the same point in time in nearby possible worlds. Given determinism, A and B are equally determined to act as they do in all possible worlds, including their act

of z'ing in $t1^{@}$. According to the counterfactual account of alternative possibilities, the fact that A, but not B, acquires basic desert in $t1^{@}$ is explained by A's, in $t1^{@}$, dispositional property (that she has in $t1^{@}$) to be x'ing in t1' and y'ing in t1*, a dispositional property that B lacks. But regardless of this, both A and B have the disposition to perform z in $t1^{@}$ and none of them, of course, has the disposition to act in any other way than they actually do in $t1^{@}$.

A's and B's dispositional properties seem to be different in terms of their *functional* nature: given identical circumstances, they will give different outputs. It is quite easy to get the intuition behind the thought that this functional difference is a relevant difference with regard to basic desert: A is the kind of person who, over a range of possible situations, responds to the relevant reasons in the situation, and thus, the idea is that A is the kind of person who *can* respond to the relevant reasons. And since A can (in the dispositional sense) respond to reasons also in $t1^{@}$, but fails to do so, it makes sense to hold her morally responsible (in the relevant sense) for her failure to respond to the relevant reasons. Consequently, according to this line of thought, A's dispositional property to respond to reasons is the base property of basic desert. However, given determinism, both A and B are equally disposed to do what they actually do in $t1^{@}$, that is, they are equally determined not to respond to the relevant reasons in that situation. In this sense, for both of them, doing something other than z in $t1^{@}$ was equally ruled out by the past and the laws of nature. At $t1^{@}$, was equally impossible for A to do y or x, as it was for B to do y or z, and it is hard to see why the fact that A in other (non-actual) possible worlds is determined to do other things, such as x and y, would be relevant to whether A, but not B, is responsible in the basic desert sense when both of them do z. When their respective abilities to act otherwise are described like this, it is not obvious why A's capacity to act otherwise under other circumstances is enough to establish that she is relevantly different from B at $t1^{@}$.

Hence, when looking at A and B doing z at $t1^{@}$, the difference in dispositional properties seems not enough for establishing that there is a *relevant* difference between A and B such that this difference can be the basis of basic desert. Because at $t1^{@}$ A and B are equally determined to act as they do, and they are equally unable to act

otherwise. Therefore, the relevant difference condition is not satisfied, and according to the consistency requirement, A and B should not have different moral properties.

If one wants to escape the challenge from determinism by arguing that the important thing with regard to moral responsibility has nothing to do with determinism, but rather, how different mental states and actions come about, we instead face the challenge from physicalism. The problem is as follows: if we accept that the relevant difference between someone who has basic desert and someone who has not is located in some property involved in the sequence of events that leads to the action in question, there must be an intuitively plausible account of what it is that makes that property special in the sense that it is the base property of basic desert. Moreover, this intuitive account must not be based on consequentialist considerations, such as Vincent's idea that a certain practice is more beneficial than burdensome for the individual person. Rather, in order to make sense of basic desert, it must be an account that makes it intuitively plausible to connect a certain natural property to the moral property of basic desert, without invoking instrumentalist concerns. In order to pin down the properties between which we must pick out this relevant difference, a plausible method could be to determine which properties in a mental event that is causally relevant with regard to an action.

Since we accept physicalism, and follow Kim (2000) regarding the exclusion problem the causally relevant property of a mental event is either identical to, reducible to, or otherwise dependent on a neural cause. Plausibly, this neural cause is a neural state in the brain.[45] Hence, the relevant difference between two mental events, where one includes the base-property of basic desert and one does not, is to be found on the neural level. But on the level of brain states, the Principle of Relevant Difference is hard to satisfy: there seems to be no relevant difference between brain states that cause the kind of actions for which we are (supposedly) morally responsible and brain states that cause other kinds of actions for which we are (supposedly) not morally responsible – they seem to be processes of the very same

[45] For Kim's view of the exclusion problem see p. 102.

kind. Hence, the challenge from physicalism makes it hard to satisfy the relevant difference condition, since the causal power of mental states are found at the level of brain states. It seems to violate the consistency requirement that one neural property, but not another, is the base-property of basic desert, if they are of the very same kind in the sense that they work in accordance with certain input-output mechanisms, and the way they process incoming input depends completely on their physical and functional properties.

The worry that the dispositional, functional properties that are picked out as base-properties for basic desert do not meet the Principle of Relevant Difference resembles what Watson pointed at as problematic for Frankfurt's hierarchical theory of free will. Watson's critique was the following:

> The problem […] is not that there is a regressive ascent up the hierarchy, or that people are not that complex, but simply that higher-order volitions are just, after all, desires, and nothing about their level gives them any special authority […] If they have that authority, they are *given* it by something else. (Watson, 1987, p. 149)

Watson's critique is, in a sense, concerned with the Principle of Relevant Difference: if higher-order volitions should play a special role with regard to free will, they must be relevantly different from first-level volitions. Otherwise, to give them a special normative authority violates the consistency requirement. If we apply a similar critique to counterfactual accounts of free will, we can ask: what is it with the dispositional properties of A that makes them special with regard to moral responsibility, when they are, in their more basic nature, not relevantly different from the dispositional properties of B? If the properties of A have that authority (with regard to moral responsibility) they are *given* it by something else. For example, we can give certain properties a certain authority with regard to desert, of e.g., consequentialist considerations, as Vincent does. But to give some natural properties such an authority does not provide us with *basic* desert, since basic desert is ascribed to a person solely based on an action that she has performed and certain intrinsic properties of her that make her into a person who can deserve things on the basis of what she has done. According to such a view, we do not ascribe basic

desert to people because it has desirable consequences, but because they *actually deserve* something, given their natural properties.

In this section I have discussed the prospects of justifying basic desert based on what I call counterfactual accounts of alternative possibilities, and more specifically, compatibilist theories that characterize the sufficient conditions for moral responsibility in terms of dispositional capacities. I have concluded that this approach to moral responsibility cannot justify basic desert attribution, because it fails to satisfy the Principle of Relevant Difference. In the next section, I shall discuss Peter Strawson's account of moral responsibility as connected to reactive attitudes, and whether this kind of approach can ease the worry about the justification of basic desert.

4.4 Wrong focus? Strawson's diplomatic account

The essay "Freedom and Resentment" (1962), P.F Strawson's famous contribution to the free will and responsibility debate, is an attempt to bring incompatibilists and compatibilists closer to each other. Instead of focusing on the possible consequences of determinism for free will and responsibility (he explicitly claims that he does not know what the thesis of determinism is (Strawson, 1962a, p. 1)) he focuses on the nature of moral responsibility. According to him both compatibilists and incompatibilists have misconstrued the nature of moral responsibility, and hence arguments from both sides miss the target. His attempt to draw a distinction between different "modes of interpretation" is similar to what some more recent thinkers, e.g. Morse (2013b) and Ruth Ann Mackor (2013), argue is the way we must understand neuroscientific findings in the context of law: even if neuroscience gives us a much richer theoretical understanding of how human behavior is connected to brain functions, brain functions are not what we talk about when we discuss for example people's desires, wishes, beliefs and so on.

In Strawson's view, responsibility judgements are connected to deep and innate features of the human psyche: it is about gratitude, resentment, forgiveness, love and hurt feelings: this emotional space is where responsibility judgements come from. We react to the

behavior of others, and care about how others react to our own behavior. Strawson makes a distinction between *participant* reactive attitudes and *objective* attitudes, where the first kind of attitudes " […] are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions" (Strawson, 1962b, p. 67). That is, the attitudes we have when we *participate* in human relationships. Our practice of holding people morally responsible is part of this practice of having reactive attitudes when we interact with people. The second kind of attitudes, objective attitudes, are such that we can choose to apply them in some circumstances. For example, a teacher working with a child diagnosed with a neuropsychiatric disorder might try to set her participant reactive attitudes aside, and adopt the objective reactive attitudes towards the difficulties of this child to concentrate, follow instructions, etc. When the child displays her difficulties, the teacher tries to identify what in the situation that has triggered the behavior and adapt the circumstances in ways that serves the child's learning process better.

When adopting the objective attitude towards another person, we see her as an object rather than a social subject with which we interact. This means that the *interactive* part of the situation disappears. The objective perspective can be accompanied by emotions like, for example anger; fear; disgust; or love; but these emotions serve another purpose in the objective attitudes compared with the participant reactive attitudes. As in other situations when we have such emotions directed *towards an object* (as opposed to having the same emotions in an interactive situation) they are not there in order to fulfill any interactive purpose.

Even if participant and objective attitudes are not mutually exclusive, they are *opposed* to each other (Strawson 1962b, p. 66). To take a *fully* objective perspective on a person seems then to exclude the possibility to have a *fully* participant perspective on the very same person. Strawson's theory is construed around the question of what effect the acceptance of the truth of a general thesis of determinism has upon participant reactive attitudes. Should the acceptance of the truth of the thesis lead to the decay or the repudiation of all such attitudes? Should we stop feeling gratitude, resentment, forgiveness,

love, and hurt feelings? Well, it is conceivable, in Strawson's view, that all these things *may* happen. But is it practically possible? Strawson believes not. The alternative to replace all participant reactive attitudes with objective reactive attitudes and start viewing each other as mere objects or automata with which we have no genuine interactive, social contact is perhaps possible in a theoretical discussion in which we discuss possible ways of treating each other from an intellectual point of view. But this move, Strawson points out, actually seems to take us away from what we actually are – interactive and social beings.

Strawson's aim was to turn the focus away from the metaphysical discussion of free will and responsibility, as, for example, the discussion of whether moral responsibility compatible with a determinist universe. But although his account of responsibility provides some intuitively plausible explanations to e.g., why responsibility is central in social interactions, the metaphysical questions remain. As was considered to some extent in chapter 3, it is hard to justify any internal practice (such as e.g., the legal proceedings) without referring to some external justification: even if we may accept that people's reactive attitudes *de facto* seem to work in a certain way, we may still ask whether the practices that evolve out of these reactive attitudes are justified with regard to other moral considerations. Further, it may be argued that even though reactive attitudes play a central role in human interaction, that does not exclude that the ability to analyze and asking for justification of those attitudes is central as well. After all, it is plausibly a characteristic feature of human psychology that we look for justification and rational support for our actions and practices.

Ayer (1962) complains of "intellectual discomfort" by Strawson's sharp demarcation between participant and objective perspectives. It is however not the case that Ayer does not agree with Strawson about the meaningfulness of distinguishing between an objective and participating perspective on people's behavior. Even if Strawson provides no argument for why the participating perspective is more valuable for us as humans than the objective perspective, Ayer expresses that he is strongly inclined to side with Strawson on this matter. What troubles Ayer is that he does not care for having

attitudes that are obviously irrational. One way to judge if an attitude is irrational is by measuring the probable consequences of adopting it: does the attitude help us to achieve what we want or need? Another way to measure the rationality of an attitude is by measuring the reasons we have for the belief to which the attitude is attached. If we consider a participating reactive attitude such as anger, we are most often angry *because of* something, i.e., we believe that something is the case. (If we do not know why we are angry, it is doubtful if the attitude can be called "reactive" since it is unclear whether the attitude is a reaction to something.) If we, on reflection, see that we have no good reasons for accepting the belief connected to the reactive attitude in question, this insight becomes a reason for us not only to discard the belief, but also the connected attitude. This is the problem that troubles Ayer in relation to Strawson's theory of free will and responsibility. Ayer maintains that the concept of desert that we normally employ is deeply problematic since it is built upon a metaphysical idea of self-determination which Strawson himself dismisses as inane (Ayer, 1962, p. 45). Then, if we have no reason to accept this belief in self-determination, the belief in desert seems unwarranted, too. And since our attitudes such as blame and resentment rest upon these beliefs in desert and self-determination, Ayer concludes, in sharp contrast to Strawson, that he should refrain from adhering such irrational attitudes as far as he can. This said, Ayer does not commit himself to treat either himself and others as machines. Like Strawson, he sees this outcome as an impoverishment of human life – but unlike Strawson, with an accompanying feeling of intellectual discomfort (Ayer 1962, p. 46).

Paul Russell (2008) also expresses slight discomfort with Strawson's view on the incompatibility of reactive attitudes and rational thinking. Russell introduces two distinctions, one between the "rationalistic" and "naturalistic" perspectives on reactive attitudes, and another between "type" and "token" pessimism concerning the aptness of reactive attitudes (Russell, 2008, p. 145).

The rationalistic strategy is displayed in Ayer's critique of Strawson's theory, where the argument relies on our rational ability to recognize that our reactive attitudes are misplaced in some, or all circumstances since they rest on unwarranted beliefs. Even if Ayer

admits that it seems very hard, if not impossible, to abandon the natural reactions to our own and others behavior, it is, nonetheless, the rational conclusion we must draw when analyzing the beliefs that ground our responsibility practices more closely. The naturalistic strategy, on the other hand, entails that it is psychologically impossible for us to entirely suspend or abandon our reactive attitudes – they are so thoroughgoing and deeply rooted in us so that they are practically insulated from sceptic doubts about their aptness. As a feature given by the human nature, claims to the effect that our reactive attitudes require external justification are mistaken, just as much as it is a mistake to ask for justification for the human ability to be afraid. Emotions do not require external, rational justification: we cannot, through reason, discover that emotional reactions are unjustified and then get rid of them. Strawson argues that since emotions is not a category that requires justification, emotions cannot be shown to be unjustified by scientific findings or rational deliberation, and whether they are determined or not does not matter for their justification, either.

According to Russell, Strawson makes a mistake here by misrepresenting the typical sceptic argument: Strawson does not recognize the difference between being a *type*-sceptic about the aptness of reactive attitudes and being a *token*-sceptic. Even if Strawson is not explicit about what he has in mind, Russell argues that it is reasonable to interpret him as if he discusses type-pessimism. But, Russell argues, this understanding of the typical sceptic argument misrepresents her actual concern. Usually the sceptic does not claim that the emotions *as such* are unjustified. It is not like anger *as such,* or resentment *as such,* are unjustified kinds of emotions. Instead, what the sceptic claims is that, in every individual case, i.e., for every *token* of these kinds of emotions, they seem to co-exist with certain beliefs, which are likely to be false beliefs. If this is the case, it means that no token of these reactive attitudes has a proper target.

Both Ayer and Russell criticize Strawson's argument to the effect that reactive attitudes are immune to reason: it does not really bite, and the reason is that attitudes, at least when interpreted as tokens, simply *are* not immune to reason. Ayer discusses the intuitively attractive idea that when we know that a certain belief most probably

is unjustified, then this insight will affect the connected attitude as well. Perhaps not in the sense that the attitude is wiped away completely, but in the sense that we are less inclined to act on it. Russell shares this view, and writes:

> [T]he sceptic finds Strawson's naturalistic reply both misguided and disturbing. What is particularly disturbing […] is that it casts doubts on our ability, or capacity to curb or control our emotional life according to the dictates of reason. More specifically, it seems clear that, despite disclaimers to the contrary, Strawson's naturalistic strategy invites us to accept or reconcile ourselves to reactive attitudes (and their associated retributive practices) even in circumstances when we have reason to repudiate them. (Russell, 2008, p. 152)

Neither Russell nor Ayer dismiss Strawson's claim that our reactive attitudes are deeply rooted parts of our human nature – but they reject the idea that this part of the human nature cannot be influenced, and changed, by rational deliberation.

I believe Russell's and Ayer's critique of Strawson's account to be instructive for the main purpose of this book: to analyze the Revision Argument and discuss whether legal retributive punishment is justified, and what is required for such a justification. As I said in the beginning of this section, I find Strawson's general approach to free will and moral responsibility useful, since we need an account of moral responsibility that people in general are inclined to accept: in other words, an account of moral responsibility that is, or has the prospects of being, supported by folk morality. Otherwise, our theory cannot be a candidate for the role of justifying legal punishment practice. However, I also agree with Russell and Ayer in their complaints that Strawson's approach misrepresents certain aspects of our responsibility practice, and especially the complex relation between reactive attitudes and rational thinking. Moreover, Strawson's theory does not take us any further concerning the metaphysical discussions of basic desert, but the dialectics between Strawson's account and the objections from Ayer and Russell are useful to think about the bigger picture: how our responsibility practices reflect both the social nature of human interaction, and that they are also proper targets of critical and rational thinking. Furthermore, it seems reasonable to agree with Strawson that reactive

attitudes have a central role in our social life, but that, in contrast to Strawson, this central role is a reason for why we should scrutinize if the moral practices that are connected to these reactive attitudes are justified.

These issues will be returned to in chapter six, where the intricate relationship between folk morality, folk psychology and science will be further discussed.

4.5 Summary & conclusions

In this chapter I have discussed the possibilities of justifying basic desert in a compatibilist framework. The background to this discussion is Morse's objection to the Revision Argument, presented in chapter 3. In the Revision Argument, it is claimed that legal retributive punishment presupposes a libertarian free will for its justification. According to Morse, legal retributive punishment presupposes neither libertarianism nor compatibilism about free will – as a matter of fact, the law does not include assumptions of free will at all. Further, if we ask for an external justification for the retributive element in the legal system, Morse claims that "only compatibilism provides a secure basis for criminal responsibility" (Morse 2013a, p. 28).

Since the legal system most plausibly needs external justification, the question in this chapter has been whether compatibilism can provide a secure basis of criminal responsibility in the sense of justifying basic desert. This question was discussed in light of a moral principle called "the Principle of Relevant Difference" and two prima facie challenges for compatibilism that arise from this principle, given that we accept the metaphysical doctrines of determinism and physicalism.

I have discussed two kinds of compatibilist theories. I called them "actual sequence compatibilism," according to which PAP is not necessary for moral responsibility, and "counterfactual theories of alternative possibilities" according to which a person has the possibility of doing otherwise in virtue of certain dispositional properties.

I argued that the challenges from determinism and physicalism provide problems with regard to the possibility for these compatibilist theories to justify basic desert.

Towards the end of the chapter, I briefly discussed if P.F Strawson's (1962) approach to moral responsibility can support Morse's criticism. According to Strawson, both incompatibilists and compatibilists have misconstrued the nature of moral responsibility. In Strawson's theory, the justification of responsibility judgements should not be analyzed in relation to determinism, but in relation to their role in social interaction. Some critique of Strawson's perspective was considered: Ayer (1962) complains that Strawson's sharp demarcation between objective and participant perspective does not make sense (or rather, that it creates "intellectual discomfort") since both these perspectives are involved in responsibility judgements. Russell (2008) put forth a similar objection: he suggests that Strawson misrepresents the typical sceptic argument. The sceptic does not claim that reactive attitudes as e.g., resentment is unjustified *as such,* but that they co-exist with a certain kind of belief, and the belief is, according to the sceptic, mistaken. It is this connection between attitudes and (unjustified) beliefs that makes responsibility judgements unjustified. In sum, I concluded that Strawson's account cannot, in itself, contribute to the discussion of how to justify basic desert as it is discussed in this chapter, that is, in the light of the principle of relevant difference and the metaphysical constraints of determinism and physicalism.

I have concluded that none of the compatibilist approaches to moral responsibility discussed in this chapter can justify basic desert attribution, since none of them satisfy the Principle of Relevant Difference in light of the challenges from determinism and physicalism. If this conclusion is correct, it means that Morse's compatibilist objection fails to make a convincing case against the Revision Argument.

In next chapter, I shall discuss a view that is a possible problem for the arguments that I have put forward in this chapter. According to this view, if we accept non-reductive physicalism rather than reductive physicalism, then it is conceptually confused to think that claims about the brain can tell us something about the mental states

upon which our responsibility attributions are based. Thus, the Revision Argument, and my arguments in this chapter, would be flawed, since they are based on a conceptual confusion.

5 Second objection: Conceptual confusion about the nature of mental states

> For Churchland, the mind is the brain and so normativity must be in the brain. For Greene & Cohen, the mind is the brain and, so, if normativity is anywhere it must be in the brain- but because they do not see it there, they conclude it is nowhere. We reject both pictures. (Pardo & Patterson, 2014, p. 42)

5.1 Introduction

According to the Revision Argument, legal retributive punishment is justified by folk psychological and folk moral accounts of free will and responsibility. It is claimed that according to these accounts, in order to deserve punishment, one must be responsible for one's actions, and in order to be responsible for one's actions one must be able to act freely, and in order to act freely, one must have a libertarian free will. But, the argument goes, if we accept determinism, no one can have a libertarian free will, since actions are determined by earlier events and the laws of nature. The view that human action works according to deterministic and mechanical principles is supported by neuroscience, since people's behavior can be accounted for in purely neuroscientific terms. In a neuroscientific explanation of action, free will in not required in order to make sense of why people act as they do. Hence, neuroscience supports the view that (libertarian) free will is not involved in human action.

In their book "Minds, Brains and Law: The Conceptual Foundations of Law and Neuroscience" (2013), Michael S. Pardo and Dennis Patterson argue that the neurolaw debate is partly shaped by conceptually confused claims about mental states. These conceptual muddles not only lead to the wrong conclusions of what neuroscience can tell us about minds and mental states, but they also lead to false

normative claims, for example about responsibility and the justification of punishment.

That brain science is of limited relevance for inquires of free will and responsibility is a well-established position in the neurolaw debate. This general approach is based on a number of different arguments that have much support among philosophers and legal scholars alike. The objection that Pardo and Patterson put forward is most notably also defended by Maxwell Bennett & Peter Hacker (2003), in their well-cited book "Philosophical Foundations of Neuroscience". In this chapter, this position will be interpreted specifically as an objection to the Revision Argument, referred to as "the Conceptual Objection".

In accordance with Morse, Pardo & Patterson claim that "one can coherently be a compatibilist and a retributivist, a combination that is consistent with current law" (Pardo & Patterson 2013, p. 198).[46] In the previous chapter, I discussed the possibility of justifying basic desert attribution, which is necessary for the justification of retributivism, within a compatibilist framework. I concluded that this approach was problematic in light of the Principle of Relevant Difference and the challenges from determinism and physicalism.

Pardo & Patterson's approach to this discussion is that the justification of retributivism, i.e., the justification of the attribution of basic desert, is intimately connected to mental state explanations. But in order to understand how this justification works, it is important to understand what mental states are. According to Pardo & Patterson, the main problem with the Revision Argument is that it rests upon an untenable account of the nature of mental states. This untenable account is based on a misunderstanding of how mental concepts work. But, they argue, if we scrutinize the nature of mental concepts, we will realize that there is a more plausible account of the nature of mental states, and that this account can do the job of justifying basic desert attribution.

---

[46] By "coherently" I take them to mean that retributivism can be justified within a compatibilist theory of free will and/or moral responsibility, since they partly frame their theory as a defence of retributive punishment against incompatibilist argument, such as the Revision Argument.

In this chapter, I will discuss my interpretation of Pardo & Patterson's argument, as divided into the following claims:

1. On a correct view of mental state concepts, mental states cannot be reduced to brain states. Therefore, mental explanations of actions cannot be reduced to explanations in terms of brain states (see Pardo & Patterson, 2013, p. 41).

2. People can be responsible and deserve punishment (in the sense of basic desert) for their actions, if their actions can be explained in terms of mental concepts such as choices (see Pardo & Patterson, 2013, p. 202).

3. From (1) and (2) it follows that brain state explanations are irrelevant to whether people are responsible in the sense of basic desert or not.

The claims above serve as an objection to the Revision Argument, in the sense that they reject the idea that neuroscientific explanations of actions can inform us about the justification of basic desert-attribution. Furthermore, they also serve as an objection to the conclusion in chapter four, that compatibilism is unable to provide a secure basis for retributivism. In that chapter, I argued that in order to justify basic desert attributions, there must be a *relevant difference at the level of brain states* between mental states that are the supervenience bases of basic desert, and mental states that are not. But according to the objection from Pardo & Patterson, this line of reasoning makes no sense. With regard to the justification of desert-attribution, the relevant differences between someone who has the moral property of basic desert and someone who has not cannot be found on the level of brain states. I will label this view "non-reductive compatibilism". According to non-reductive compatibilism we cannot look for the justification of responsibility and desert at the physical level, since the mental states in virtue of which people are responsible and have basic desert, cannot be reduced to physical states. In this chapter I will discuss whether non-reductive compatibilism is a viable view, and whether it can be used to refute the Revision Argument and the

arguments against compatibilism about basic desert presented in chapters three and four.

This chapter will run as follows: In section 5.2, I will describe Pardo & Patterson's claim that mental concepts cannot be reduced to physical concepts. In section 5.3, I will describe why, according to Pardo & Patterson, choices are the basis of responsibility. In section 5.4, I suggest that their view of mental states is best construed as a functionalist theory. One worry with a functionalist theory of mental states is that it is difficult to make sense of mental causation, and mental causation is central for the explanatory relevance of mental states in relation to actions. In section 5.5, I will address this worry, and (briefly) discuss two non-reductive approaches to mental causation, in order to analyze whether Pardo & Patterson's functionalist view of mental states can be combined with a view of mental causation that can satisfy the Principle of Relevant Difference in light of the challenges of determinism and physicalism. The first view I will discuss is that of Donald Davidson (1970), according to whom we can keep a monistic physicalist theory of the mental, without giving up non-reduction about mental states and mental causation. The second view I will discuss is a more recent contribution to the debate of how to conceive of mental causation: Christian List & Peter Menzies (2017) suggest that mental events fulfill two criteria that they claim are central when we think of causation, and that both mental and physical events can be complete causes (however, not in the same sense of "cause"), of an action. I will argue that neither of these views show convincingly why we should not look at the level of brain processes when we look for the relevant difference between natural properties (e.g., reasons-responsiveness) that are the base-property of basic desert, and natural properties that are not, and that they therefore face the same requirements from the Principle of Relevant Difference in light of the challenges from determinism and physicalism as the compatibilist theories that were discussed in chapters three and four. I will conclude that the Conceptual Argument does not refute the Revision Argument, or the arguments that I advanced against the counterfactual theories of alternative possibilities in chapter three and four. A summary of the chapter is provided in section 5.6.

5.2 The irreducibility of mental concepts and mental states

The Conceptual Objection is framed as a critique against the Revision Argument. The central claim in the Conceptual Objection is that mental states cannot be reduced to brain states. Therefore, mental explanations of actions cannot be reduced to explanations of actions in terms of brain states.

A clear understanding of how neuroscientific research contributes to our understanding of mental states, as well as its potential legal application, "requires a clear articulation of psychological concepts presupposed by it [neuroscience] and the psychological capacities about which the research seeks to provide empirical evidence" (Pardo & Patterson, 2013, p. 96). We need to be careful with the distinction between the claims "the mind *depends* on the brain" and "the mind *is* the brain" (p. xiv). To say that the mind depends on the brain implies that the brain in some sense is necessary for the mind, but does not specify in what sense. The statement that the mind *is* the brain points at a reduction-relation: i.e., one thing is reducible to the other. According to this claim, minds are not something separate from brains. Hence, in order to understand the nature of minds, we must understand the nature of brains.

Pardo & Patterson explicitly reject this reductionist approach to mental states (2013, pp. xiii, footnote 2). To illustrate why, they point out several empirical and conceptual issues that need to be considered in order to understand the relation between minds and brains. The empirical issues concern e.g., the adequacy of the scientific explanations for the phenomena that we associate with the mind. The conceptual issues are sometimes more difficult to recognize, both with regard to the general relationship between minds and brains, but also with regard to specific mental categories. For example, in the claim "depression is a certain brain chemistry state" it is presupposed that the mental state of depression is "located" somewhere and that we can "find" this mental state and examine it through brain science. But if, in fact, the term *depression* does not refer to a brain chemistry state but to certain psychological states, such as certain experiences of sadness, and, Pardo & Patterson argue, to certain actions that are

connected to these psychological states, it makes no sense to say that depression is a certain brain chemistry state (Pardo & Patterson, 2013, p. 113). They admit that we can change the way we use the concept of depression, and start referring to a certain brain chemistry state instead of a psychological state and behavior. But changing the concept of depression does not change the fact that when this particular brain chemistry state is actualized, people will experience a psychological state of sadness (and behave in accordance with this.) And this psychological state, in turn, cannot be explained in terms of a brain chemistry state alone: it is not an analytical truth that a certain brain chemistry state corresponds to the psychological state of sadness and certain kinds of behavior that is connected to this psychological state. Pardo & Patterson argue that mental states cannot be explained in terms of brain states because mental concepts denote a wide range of phenomena, most of them behavioral. One example is the mental state of believing something:

> For belief, this includes, for example, that one asserts or endorses what one believes, that one acts in ways consistent with one's believes, that one does not believe directly contradictory propositions, and so on. This behavior is not only a way to determine whether someone perceives or believe something in particular. The behavior also helps to determine (indeed, it partly constitutes) what it means to engage in these activities. In other words, it helps to provide the measure for whether someone is in fact engaged in this activity (not just a measurement in a particular instance). (Pardo & Patterson 2013, p. 9)

According to this view, what it is to have the mental state of believing is much more than having a certain brain state. This point can be illustrated when we ask for evidence for if someone does or does not have a certain belief. Pardo & Patterson point out that we must distinguish between *criterial* and *inductive* evidence for mental states. They state:

> Criterial evidence for the ascriptions of psychological predicates, such as "to perceive" or "to believe", consists in various types of behavior. Behaving in certain ways is logically good evidence and, thus, partly constitutive of these concepts [...] for belief, this includes, for example, that one asserts or endorses what one believes, that one acts in ways consistent with one's beliefs, that one does not believe directly

> contradictory propositions, and so on. This behavior is not only a way to determine whether someone [...] believes something in particular. The behavior also helps to determine (it partly *constitutes*) what it means to engage in these activities. In other words, it helps to provide the *measure* for whether someone is in fact engaged [in believing something] (not just a measurement in a particular instance.) (Pardo & Patterson, 2010, pp. 1222-1223)

Behavioral evidence for beliefs is, of course, defeasible: people can assert propositions that they do not believe, and believe things they never explicitly assert or act upon. The primary point in the quote above, however, is that behavior is not only inductive but also criterial evidence for mental states in the sense that it partially determines what it means to have e.g., a belief. By contrast, some evidence provides only inductive support for whether someone believes something. Such inductive evidence could, for example, be a correlation between certain neural phenomena and the mental state of believing. But, Pardo & Patterson claim, "this inductive correlation only works once we know what to correlate the neural activity *with"* (Pardo & Patterson 2010, p. 1224).

Another example relevant to the legal context is the phenomenon of lying. Even though it is an empirical question if a certain person is lying on a certain occasion, and also whether a particular brain activity is correlated with lying. But in order to investigate these issues, we must know what constitutes a lie – and this is a *conceptual* question:

> The criteria for telling a lie or engaging in deception involve behavior, not neurological states. To lie requires, among other things, a false assertation (or one that the speaker believes to be false.) Roughly, deception involves believing things to be so and saying or implying the opposite, and it involves judgements about the beliefs and knowledge of the audience. At most, neuroscientific evidence might be able to provide well- grounded empirical correlations between this type of behavior and brain states. This will be *inductive* evidence. The neuroscience, in other words, may be able to provide a *measurement* regarding lies or deception, but not the *measure* of it. It is a conceptual mistake to conclude that lies take place in the brain; that a particular area of the brain chooses to lie; that neuroscience can reveal lies being "produced" in the brain; or that it can "peer into" one's brain and see the thoughts that constitute lies or deception. (Pardo & Patterson, 2013, pp. 100-101)

Pardo & Patterson argue that if we acknowledge the difference between criterial and inductive evidence for mental states, it is clear why mental states cannot be reduced to brain states.

Further, Pardo & Patterson also point out that we should be aware of the hidden assumption in the claim "the mind is the brain": in order to make sense of the "is," we must think of the brain as a substance. This assumption is, they argue, a basic one in a reductive physicalist framework – but, as it were, also in the contemporarily widely rejected Cartesian dualist framework:

> The Cartesian view relies upon a notion of *substance dualism*. Under this conception, the mind is thought to be some type of non-material (i.e., non-physical) entity or thing that is part of the human being and is somehow in causally interaction with the person's body. The non-material substance that constitutes the mind is the source and location of the person's mental life – her thoughts, beliefs, sensations, and conscious experiences. Early neuroscientists were avowed Cartesian dualists and set themselves the task of figuring out how the non-material substance known as the mind causally interacted with the physical brain and body of a person. This conception was later repudiated by neuroscientists and is typically disavowed by neuro-legalists. The second conception of mind is that the mind is identical with the brain. This is the conception endorsed by Churchland, Greene, Cohen, and other neuro-legalists. Under this conception, the mind is a material (i.e., physical) part of the human being – the brain – that is distinct from, but is in causal interaction with, the rest of the body. The brain is the *subject* of the person's mental properties (the brain thinks, feels, intends, and knows) and is the *location* of the person's conscious experiences. This conception is deeply problematic. (Pardo & Patterson 2013, pp. 43-44.)

In the quote above, Pardo & Patterson point out that a problem for the Cartesian dualists, in relation to a physicalist world view, is that they believe that the mind is a non-material substance or entity of some sort, and this way of thinking made things difficult for the early neuroscientists, because they tried to find causal relations between the immaterial substance of the mind, and the brain and body. However, according to a physicalist world-view, non-physical substances cannot have causal impacts on physical entities. Later, neuroscientists gave up on the idea that there is a relation between two substances – the immaterial mind and the material body, but, Pardo & Patterson claim,

they still entertain the idea that the mind is a substance in their explanatory framework of the mind. Further, the argue that in contemporary neuroscience, the idea that the mind is a substance has led to the idea that "the mind is the brain" and by that, people like e.g., Churchland and Greene & Cohen believe that we have an answer to the questions of *what* and *where* the mind is. This means, according to Pardo & Patterson, that the basic idea from Cartesian dualism about the mind is preserved: namely, that the mind is a substance. But, they argue, this does not make any sense. If the mind is the brain, does that mean that we can attribute mental predicates to the brain? If one wishfully thinks of being on a sunny beach sipping on a cold drink, is this wishful thinking something that can be found in one's brain? The idea is absurd, they claim: the brain, in interaction with other aspects of the nervous system and the rest of the human body, makes this mental state possible. But there is no wishful thinking to be found in the brain. Because even if neural activity is necessary for us to engage in such mental exercises as wishful thinking, neural activity alone is not sufficient for successful employment of the mental concept of wishful thinking. What Pardo and Patterson think of as a more tenable view of the mind is illustrated in the following passage:

> To have a mind is to possess an array of rational and emotional powers, capacities, and abilities exhibited in thought, feeling, and action… under this conception, the mind is not a separate part of the person that causally interacts with the person's body. It is simply mental powers, abilities and capacities possessed by humans. Likewise, the ability to see is not a part of the eye that interacts with other parts of the physical eye, the ability to fly is not a separate part of an airplane, and a car's horsepower is not a separate part of the car in causal interaction with its engine. Under this conception, the question of the mind's location in the body makes no sense just as the location of eyesight within the eye makes no sense. (Pardo & Patterson, 2013, p. 44)

As Pardo and Patterson acknowledge, the view of the mental that they describe above resembles of the Aristotelian view of the mind. It is, hence, not a new way to think of the nature of mental states. The Aristotelian view played a profound role for the way European

philosophers and scientists thought of the human nature, the mental parts included, until the scientific revolution in the 17th century, when the Cartesian view of the mental replaced the Aristotelian (Pardo & Patterson 2014, p. 44). Pardo & Patterson point out that within the Cartesian framework of the mind, the question of how to give a coherent account of the mental and the physical arises quite naturally since it is presupposed that the mental and the physical are two distinct substances. The reductive physicalist still entertains the idea that the mind is a substance, even if she disagrees with the dualist that this substance is not separate from the physical body. Aristotle himself expresses his view of the relation between soul and body as follows:

> […] we should not inquire whether the soul and the body are one, any more than the wax and the shape, or, in general, the matter of a given thing and that of which it is a matter […] if the eye were an animal, sight would be its soul. And the eye is matter for sight, and if this fails, it is no longer an eye, except homonymously, like an eye in stone or in a picture. (Aristotle, 2017, p. 22)

If we paraphrase Aristotle's claim in a more contemporary wording, the unnecessary question is the one of whether the mental and the physical are identical. According to Pardo & Patterson, this point is where reductive physicalism goes astray. People who conceive of the mind as a substance may also come to think that the mind is located somewhere. But, in Pardo & Patterson's view, physicalism about the mind is fully compatible with the view that mental states are *not* located at any certain point in the body (such as, for example, in the brain). To think so is a mistake called "the mereological fallacy" (Pardo & Patterson, 2013, pp. 20-21). The mereological fallacy, thus, consists in attributing an ability or function to a part that is only properly attributable to the whole. In this specific case, it follows from the mereological principle that mental states can properly be attributed to human beings, but not to certain parts of them (e.g., to their brain). For example, claiming that one *knows* something, according to this view, is not to say anything about one's brain. Knowledge cannot be found in the brain, since knowledge is an ability: *people* have knowledge, brains do not. Knowledge, like other

psychological attributes, is essentially manifested through behavior. Only humans and animals can engage in behavior. Brains cannot behave, and behavior cannot be reduced to a particular neural event. M. R Bennett & P.M.S Hacker explain it in the following manner:

> A person who knows where the railway station is, what time the next train is, whether it is likely to be on time, who else might be on it, etc. can answer the corresponding questions. But there is no such thing as the brain knowing when…what… whether…etc., and there is no such thing as the brain's answering these questions. It is not the brain, but the person whose brain it is, that acquires knowledge by perception, reasoning and testimony. (Bennett & Hacker, 2003, p. 152)

It is also confused, according to this view, to speak of the brain as *containing* or *possessing* information. The brain is not containing information in the same way as e.g. books do:

> It is equally confused to speak […] of the brain's containing knowledge and information, which is encoded in the brain… we may say of a book that it contains all the knowledge of a lifetime's work of a scholar, or of a filing cabinet that it contains all the available knowledge, duly card-indexed, about Julius Caesar. This means that the pages of the book or the cards in the filing cabinet have written in them *expressions* of a large number of known truths. In this sense, the brain *contains* no knowledge whatsoever. There are no symbols in the brain that by their array express a single proposition, let alone a proposition that is known to be true. Of course, in this sense a human being *contains* no knowledge either. To possess knowledge is not to contain knowledge. A person may possess, for example, a smattering knowledge about seventeenth-century woodcuts, but his brain contains none…the brain neither possesses nor contains any knowledge. (Bennett & Hacker, 2003, p. 152)

According to the view displayed in the above examples, mental states, as for example the mental state of knowing, is not "in" the person, but it is *displayed* in certain ways by the person. If we accept this view of mental states, it seems at that we can gain limited information about mental states through neuroscientific studies of the brain.

In sum, Pardo & Patterson's understanding of mental states is compatible with physicalism, which is a presupposition in the discussions in this thesis, but not with reductive physicalism (which is not a presupposition in this thesis.) The relationship between the

brain and the mind is, according to Pardo & Patterson, roughly that brain states are inductive evidence of the presence of a mental state, but they are not criterial evidence of mental states. Mental states are not substances, and they are not possible to locate at a special point in the body. To say that someone has e.g., a belief is, in Pardo & Patterson's view, not to say that some part of her brain works in a certain way: rather, it is more like a description of her that is correct if she fulfills certain evidential criteria. In the next section, I will discuss their view of responsibility and basic desert. They argue that explanations of responsibility and basic desert must be connected to mental states, and since mental states are irreducible, we cannot investigate the basis of such practices on the level of brain states.

5.2.1 Choices and actions as the basis of responsibility (and basic desert)

> We have a choice. Is it this choice that is the ground of responsibility, which cannot be accounted for in eliminativist terms. (Pardo & Patterson, 2013, p. 41)

Similar with Morse, Pardo & Patterson think that libertarian free will is not required as justification for retributive punishment:

> Sufficient control over one's actions in light if one's practical rationality is sufficient to ground moral desert, regardless of whether the same actions may be explained in purely physical (i.e., nonmental) terms. In other words, one can coherently be a compatibilist and a retributivist, a combination that is consistent with current law…[t]he idea that people possess the opportunity to do otherwise is consistent with determinism. (2013, p. 198-99)

Having "sufficient control" means, according to Pardo & Patterson, "the presence of both an ability and an opportunity to exercise it [...] agents who have the ability and the opportunity to act differently, but do not, are properly subject to moral evaluation" (Pardo & Patterson 2013, p. 199). In their view, this way of thinking makes sense in the light of what makes it true that someone has an ability and an opportunity to act differently:

> To possess an ability (e.g., to ride a bicycle) depends on whether one satisfies the criteria for possessing the ability. These criteria include successfully exercising this ability when one wants to (and has the opportunity to) and refraining when one wants to (and has the opportunity to refrain). Such criteria can be fulfilled even if one does not exercise the ability on the particular occasion in question. (2013, p. 199)

To have an ability is, in Pardo & Patterson's view, at least partly a matter of being able to exercise that ability when one wants to. This view is similar to the counterfactual theories of alternate possibilities discussed in chapter four, according to which abilities are dispositional properties: if someone has the ability to ride a bicycle, she can do so if she wants to (and the circumstances allow for it.) Moreover, Pardo & Patterson point out that there is an important difference between the claim that brain states involved in our actions, and the claim that our brain states "force" us to act as we do:

> […] are an agent's brain states forcing him to act in one way and preventing him from acting in another (are they an "external force rigging his behavior") and, thus, depriving him of the opportunity to do otherwise? Not necessarily. We presume that if the agent had wanted to do something different (e.g., to ride a bicycle or not) then his brain states also would have been different. It would have been a different story if his brain states caused him to ride a bicycle (or not) when he wanted to do the contrary. In such circumstances, there would be a breakdown of the type of rational control on which criminal responsibility depends. (Pardo & Patterson 2013, pp. 199- 200)

Even if Pardo & Patterson do not explicitly discuss reasons-responsiveness, they point at the dispositional ability to do otherwise as the basis of basic desert. When these issues were discussed in the previous chapter, they were related to the metaphysical constraints of physicalism and determinism, in combination with the Principle of Relevant Difference. I argued that in order for it to be intuitively plausible for reasons-responsiveness to be the supervenience-base of the moral property of basic desert, there must be a relevant difference between reasons-responsiveness and other natural properties that, intuitively, are not supervenience-bases of basic desert. The Principle of Relevant Difference is a comparative principle, saying that there must be a relevant difference in natural properties between two

actions, events, or agents in order for them to differ with regard to moral properties. For basic desert, this principle means that *ceteris paribus,* there must be a relevant difference in natural, mental properties in order to justify the moral difference between P and Q such that P has the moral property of basic desert and Q does not. But in an actual situation where P and Q perform identical actions and P is reasons-responsive and Q is not, P's *dispositional* property of being reasons-responsive is not involved in the *actual* processes that lead to the action under consideration, and it seems difficult to pick out a difference between the actual mental processes involved in the causal process that leads to P's and Q's action respectively, such that this difference can justify the moral difference between them (i.e., the difference that P has basic desert and Q does not.)

However, this analysis is built upon the supposition that the role of reasons-responsiveness for basic desert (or any other dispositional property that is sufficient for basic desert) has to do with its *causal connection* to actions, and Pardo & Patterson questions this presumption. Not in the sense that causal explanations are irrelevant to explanations of behavior, but because explanations that describe causal relations in terms of mental events and actions cannot be reduced to explanations that describes causal relations in terms of physical causes and effects. Consider, for example, the question of why a person stops her car at a red light. This behavior, Pardo & Patterson contend, has to be explained appealing to a traffic rule. In a sense, we want to say that what caused the person to stop was the red lamp. But by itself, the red light does not "cause" the person to stop (i.e., it is not in virtue of the power of the light waves that emanate from it.) Rather, she stops because of the status of the light in an important social convention (i.e., the red light is a reason for stopping) (Pardo & Patterson, 2013, pp. 40-41). In this explanation, Pardo and Patterson argue, we look for what kind of reasons that were considered when the person decided to stop the car, not what causal, mechanical processes that lead to the pressure of the driver's foot on the brake.

As I argued in the previous chapter, the central role of mental causation in basic desert-attribution forces us to look for the relevant differences between a person who has the moral property of basic

desert and a person who does not on the level of physical (brain) properties. The reason was that if basic desert is ascribed to people on the basis of their mental properties, it seems intuitively plausible that mental properties that are base properties of basic desert have some kind of causal effect with regard to how actions come about, that makes a difference compared to how action come about without the presence of this mental property. But in Pardo & Patterson's view, this reasoning is mistaken. They argue the causal role of a certain mental capacity in relation to actions cannot be equated with physical causal relations: the "because of" in a mental explanation of an action cannot be reduced to a "because of" in a physical explanation. They put it as follows:

> When a bowling ball hits the pins, we say that the pins fell over because they were hit by the ball. The reason the pins fell over is that they were hit by a bowling ball. One event – the pins falling over – was caused by another – the bowling ball hitting the pins [...] Unlike the pins, we choose whether to stop at the red light. If we fail to stop, we run the risk of sanction. The pins in a bowling alley have no such choice: they are "compelled" to fall over by force of the impact of a bowling ball. We are neither bowling balls nor pins. We have a choice. It is this choice that is the ground of responsibility, which cannot be accounted for in eliminativist terms. (Pardo & Patterson, 2013, pp. 40-41.)

In Pardo & Patterson's view, there are some fundamental differences in explanations of actions, and explanations of mechanical events such as when a bowling ball hits some bowling pins. A person's action is due to several factors, such as in the example with the traffic light. It is not sufficient to explain the causal process involved in the pressure of the driver's foot on the brake – we must also consider the social convention of traffic lights, the driver's ability to follow the traffic rules, etc. Here, the mental states of the driver play a central role if we are to understand the relation between her choices, her other mental states, and her actions. As mentioned above, someone is responsible for her actions, in Pardo & Patterson's view, if she has a kind of rational control over her behavior: they say that "if his brain states caused him to ride a bicycle (or not) when he wanted to do the contrary [...] there would be a breakdown of the type of rational control on which criminal responsibility depends"(2013, p. 200).

What distinguishes Pardo & Patterson's account from the reductive physicalist account when it comes to explanations of actions is not the idea that we need to take many different aspects into account if we want to understand why someone acted as she did. Rather, the disagreement lies in how to think about mental states that are relevant in action explanations, such as e.g., choices. According to the reductive physicalist, the actual choice someone makes in a specific situation can be reduced to – and explained by– physical mechanisms in the brain. The causal mechanisms involved when we make choices are not different from other causal relations in the world: there is not "special" kind of causation involved. The mechanisms involved when we make choices are governed by natural laws and they are, as is a presumption in this discussion, determined to their nature. Pardo & Patterson argue that choices cannot be reduced in this way. Choices are not brain states. They point out that it is true that we must have a brain in order to make choices, but this does not imply that a choice is a brain state, or that choices can be located in the brain. Therefore, it does not make sense to say that neuroscientific explanations of actions can show that people do not "really" have a choice: choices are not such things that can be identified through brain science.

In order to evaluate whether this non-reductionist approach is a powerful objection to the Revision Argument, and also an objection to my argument against the claim that reasons-responsiveness (or similar compatibilist-friendly properties) can justify basic desert attribution, we need a more detailed picture of in virtue of what some people acquire basic desert, and why other people do not. In the next section, I will elaborate on how Pardo & Patterson's account can be interpreted as a functionalist theory of mental states, and discuss whether such an approach can succeed in evading the worries I raised in the previous chapter, against the possibility for compatibilism to be a solid ground for basic desert.

5.3 The Conceptual Objection & functionalism

In this section I will suggest that Pardo & Patterson's approach to mental states can be interpreted as a functionalist theory, and discuss

whether functionalism about mental states can provide a secure basis for the practice of legal retributive punishment by providing a non-reductivist compatibilist account of basic desert. To be successful, the account must provide a plausible explanation of why we are justified in attributing basic desert when someone chooses to act in a certain way, and what the relevant difference is between a free action and a non-free action with regard to desert attribution. This discussion will, in line with the discussion in the previous chapter, be related to the metaphysical constraints of physicalism and determinism and the Principle of Relevant Difference.

In a more general context, a theory of mental states aims to explain how mental phenomena fit into one's overall ontology. For example, the theory is supposed to give an account of the nature of mental states, and how mental states are related to other objects and phenomena in one's general metaphysics. One central question for any theory of mental states is how mental causation is possible (see e.g., Mackie 1979, Shoemaker 2001). We often explain people's actions as causally related to mental states. For example, I just went to the fridge because I was thirsty, and wanted some cold water that I believed was in the fridge. Moreover, mental state explanations seem to be relevant also with regard to basic desert. For example, it seems plausible to attribute basic desert to someone who intentionally hurts another person, but not to someone who hurts another person accidentally. For our moral intuitions concerned with basic desert, it matters if an action is carried out by purpose or if it is an accident. Philosophers such as Donald Davidson (1963) and Alfred Mele (1992) have argued that if mental states were causally isolated from bodily behavior, then what goes on in your mind cannot explain what you do, and then it seems less plausible that mental states matter to basic desert attribution.

Being a physicalist, one's theory of the mental should ideally provide an explanation of how mental states can be causally effective in a physical world. As was stated in section 1.3, I follow Jackson's definition of physicalism:

> Physicalism […] claims that a complete account of what our world is like, its nature, (or, on some versions, a complete account of everything contingent about our world), can in principle be told in terms of a

> relatively small set of favoured particles, properties, and relations, the 'physical' ones. (Jackson, 1998, p. 6)

However, this kind of physicalism does not exclude the existence of non-physical phenomena. In order to account for them, Jackson introduces supervenience, according to which any non-physical entity supervenes on a physical entity, and the non-physical entity cannot change if the physical entity it supervenes upon does not change. If mental phenomena supervene on physical phenomena, any change in mental phenomena requires a change in physical phenomena. Moreover, Jackson argues that any two worlds that are physical duplicates, are also duplicates with regard to non-physical phenomena: "any world that is a minimal physical duplicate of our world is a duplicate *simpliciter* of our world" (Jackson, 1998, p.12). Furthermore, I follow Papineau's view of causal closure which implies that mental causes are physical causes. (However, I will briefly discuss an alternative view of causation in a later section of this chapter.)

Pardo & Patterson point to *choices* as playing a central role in our responsibility practice. According to the reductive physicalist, a choice can be reduced to brain states, something that Pardo & Patterson deny. However, as pointed out by e.g., Davidson and Mele, if choices play special role in our responsibility practice, they must have some sort of causal effect on our behavior, otherwise it is hard to see why choices play that special role. Pardo & Patterson claim that mental states, and especially choices, are explanatory relevant in explanations of actions and for responsibility attribution, but since mental states cannot be reduced to brain states, the causal relationship between mental states and actions cannot be understood along the same lines as "ordinary" physical causation: "The reason we stop our car at a red light cannot be explained in the manner of a bowling ball striking a set of pins" (Pardo & Patterson, 2013, p. 41).

Pardo & Patterson's view of the mind as an array of rational and emotional powers, capacities, and abilities exhibited in thought, feeling, and action, goes well together with the idea that mental states should be identified with their functions, rather than with physical states. A brief overview of functionalism was provided in section 1.5.

The form of functionalism that I take to be the most helpful and plausible for the purpose of interpreting Pardo & Patterson's non-reductive compatibilism is analytical functionalism, as discussed by for example Lewis (1972) and Armstrong (1968, 1981). This version of functionalism claims that mental terms and concepts can be translated into functional descriptions.

An important distinction in the functionalist framework is between *role-functionalism* and *realizer-functionalism* (McLaughlin, 2006). We might think of e.g., the mental state of pain as something that makes us move away from the thing causing the pain, by producing anxiety and a desire to get rid of the pain, and we can further assume that the functional role of pain in humans is realized by C-fibers firing. But it seems that we have two levels of description in this picture: (i) the higher-order property of having the relevant functional role, and (ii) the lower-level physical realizer of that functional role, in this case the C-fibers firing. According to role-functionalists, pain is identified by the higher-level relational property that is accessible to the subject. But according to realizer-functionalism, a functionalist theory of mind provides definite descriptions of whichever lower-level properties that satisfy the functional characterizations. In this view, if the physical property that occupies the functional role of pain is C-fibers firing, then pain in humans is identical with C-fibers firing.

Non-reductive compatibilism is most plausibly interpreted as a role-functionalist approach. Even though Pardo & Patterson recognize that there are neural correlates to mental states which can be discovered by empirical investigation, they claim that it is important to recognize that the empirical evidence will be concerned with *correlations* between mental states and neural states. To think otherwise, e.g., to think that the neural mechanisms are part of the mental state itself, would be a conceptual confusion. However, as already noted, Pardo & Patterson does not they deny that brain states are involved in mental states and behavior:

> […] certain neural activity may be necessary to engage in (and play the causal role in) the behavior that constitutes the ability to think or perceive […]. (Pardo & Patterson, 2013, p. 11, footnote 34)

Neural activity is, hence, necessary in the sense that without neural activity, at least humans cannot have any mental states at all. But neural activity is not a sufficient condition for determining if someone has a specific mental state, or a specific mental capacity, since neural activity it is not included in the criterial evidence for having a certain mental state. Notably, Pardo & Patterson point out, in the above quote, that neural activity, although not sufficient for the ascription of a mental states and capacities, may play the *causal role* with regard to behavior.

Functionalism, as a general theory of mental states, remains silent about how functional states are realized in the world. Mental states might, for example, be realized by something immaterial. But in the context of this discussion, physicalism is a constraint for this theory, since Pardo & Patterson are physicalists. Hence, insofar as mental states have causal power, this power must be due to something physical. One of the main advantages of functionalism is that it allows for multiple realizability of mental states. Multiple realizability means that the same type of mental state can be realized by many different physical states, which means that we can ascribe e.g., fear to organisms with a nervous system very different from ours, which is an advantage for the theory since it seems plausible that two organisms with disparate neural systems can both experience fear. Also, different organisms can behave in different ways when they experience fear. Some organisms will be aggressive; others will be more prone to flee from the source of the fear. Functionalism can account for both of these aspects of multiple realizability in the sense that fear is defined as e.g., a stressful experience and a tendency to get out of the situation, either by being aggressive toward the feared object/organism, or by fleeing from it.

Regarding the causal power of mental states, then, if we accept both functionalism and physicalism like Pardo & Patterson, the causal power of different tokens of the same kind of mental state can, due to multiple realizability, be due to different physical states. However, Pardo & Patterson argue that in relation to responsibility attribution, it is the *intentional element* of the action that is of explanatory relevance, and in order to explain intentional action, we must refer to *reasons*. However, they also note that "[…] no explanation of human action

is complete without an account of the role of cause in behavior" (Pardo & Patterson, 2013, p.35). It is unclear whether they assume that reasons are causes, or if reasons and causes are different categories in some sense. In the footnote to the just quoted sentence, Pardo & Patterson describe their view of the causal relation between mental events and brains states as follows:

> […] there is a substantial disagreement about whether and how to best characterize the causal relationship between mental events and behavior […] These philosophical controversies are outside our scope – however one comes down on them, the conclusion regarding neuro-reductionism is the same. (Pardo & Patterson 2013, footnote 61, p. 35)

I take this last claim to be unconvincing, at least in one interpretation of it. Pardo and Patterson write that regardless of what idea we have about the role of causes in an explanation of human behavior we will conclude that non-reductionism is more plausible than neuro-reductionism. But it is not obvious why this conclusion would be reached. An important part of Pardo & Patterson's project is to convince us that neuro-reductionism is a conceptually confused theory, by arguing for a view of the mental very similar to analytical functionalism. But, as noted earlier, one of the central questions for any theory about mental states, at least given a general physicalist ontology, is how to make sense of the causal relations between the mental and the physical, especially with regard to how mental states can have causal impact on behavior. This worry is especially warranted when it comes to role- functionalism. One of the main advantages of neuro-reductionism is that it provides us with a neat account of the causal relationship between mental states and behavior, because mental states according to this theory cause behavior in virtue of being brain states, or at least, their causal power is due to brain states. Pardo & Patterson object to this view arguing that neuro-reductionism presupposes that mental states are substances, but, in their view, mental states are not. When saying that someone has a certain mental state, such as a belief, this is to say that she e.g., behaves in certain way given certain circumstances. But as such, the mental state is not located somewhere in the person's brain, in the way that neuro-reductionism presupposes. Therefore, mental

states cannot be reduced to physical processes (since physical processes are always located somewhere).

However, in order for Pardo & Patterson's characterization of mental states to be a plausible alternative to neuro-reductionism, and in particular a theory that can be used to justify retributive punishment, they must be able to explain how mental states are causally related to behavior. In the next sub-sections, I will (briefly) discuss two non-reductive approaches to mental causation, in order to analyze whether Pardo & Patterson's functionalist view of mental states can be combined with a view of mental causation that can meet the requirements of the Principle of Relevant Difference, in the light of the challenges from determinism and physicalism. If it turns out that their theory meets these challenges, and satisfies the relevant difference condition, I take their objection against the Revision Argument to be valid.

The first view I will discuss is that of Donald Davidson (1970), according to whom we can keep a monistic physicalist theory of the mental, without giving up non-reduction about mental states and mental causation. The second view I will discuss is a more recent contribution to the debate of how to conceive of causation: Christian List & Peter Menzies (2017) suggest that mental events fulfill two criteria that they claim to be central for how we think of causation and that both mental and physical events can be complete causes (however, not in the same sense of "cause") of an action.

5.4 Non-reductive physicalism & mental causation

5.4.1 Anomalous monism and mental causation

According to Pardo & Patterson, the causal explanation of actions can be found in the neural activity that is a necessary, but not sufficient, condition for mental abilities. Moreover, they claim that neural activity provides *inductive* (as opposed to criterial) evidence for that someone has a certain mental state or a mental ability. It is inductive in the sense that the criteria for if someone has a certain mental ability, e.g., the ability to lie, are behavioral in nature, and we

cannot say that someone is lying solely on the basis of certain neural activity: "this inductive correlation only works once we know what to correlate the neural activity *with"* (Pardo & Patterson, 2010, p. 1224). Even though this reasoning is intuitively plausible in a sense, for example, it seems plausible that the mental ability to lie can correspond to multiple physical states, it is unclear how mental states *qua* mental states are causally effective: if mental states are not substances, and cannot be located anywhere – how can they have causal power?

One theory (of many) that has provided an account of how to think of the causal relationship between the mental and the physical without reducing mental events to physical events is Donald Davidson's "anomalous monism", a theory about the relationship between the mind and the body that not only aims to save ontological monism, but also to make sense of mental causation without being committed to the view that mental states can be described in physical terms. The main thesis of anomalous monism is described by Davidson as follows:

> Anomalous monism resembles materialism in its claim that all events are physical, but rejects the thesis, usually considered essential to materialism, that mental phenomena can be given purely physical explanations. (Davidson, 1970, p. 119)

Davidson makes two important moves in the quote above: (1) he ascribes to physical monism, i.e., he accepts only one substance in his ontology, and (2) he maintains that even though all events are physical, mental phenomena cannot be given a purely physical explanation. The reason for (2) is not that mental events are not physical, but that the relation that holds between the mental and the physical is one of *token-identity*. In Davidson's theory, the objects between which the identity relation holds are found at the linguistic level. Despite there being just one ontological category, there are mental events and physical events at the linguistic level.[47] But what is

[47] Davidson have defended different views of how to think of the criteria of event-individuation over time: first, he endorsed a causal criterion of event individuation according to which two events are identical if they share all their causes and effects (Davidson, 1969) but he later rejected that view in favor of one according to which

a mental event if it is only found at the linguistic level? This question targets the anomalous character of the mental, as Davidson describes it: "The principle of the anomalism of the mental concerns events described as mental, for events are mental only as described" (Davidson, 1970, p. 119).

In order for an event to be a *mental* event, it must answer to mental predicates, in the same way as a physical event are such events that answers to physical predicates. Mental predicates function differently compared to physical predicates. One important feature of mental predicates is, according to Davidson, that they are defined at least partly in terms of other mental predicates. He takes the example of belief to illustrate this point:

> Suppose that we try to say, not using any mental concepts, what it is for a man to believe that there is life on Mars. One line we could take is this: when a certain sound is produced in the man's presence ("Is there life on Mars?") he produces another ("Yes"). But of course, this shows he believes there is life on Mars only if he understands English, his production of the sound was intentional, and was a response to the sounds as meaning something in English, and so on. For each discovered deficiency, we add a new provisio. Yet no matter how we patch and fit the nonmental conditions, we always find the need for an additional condition (provided he *noticed, understands,* etc.) that is mental in character… [j]ust as we cannot intelligibly assign length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions. (Davidson, 1970, pp. 120-122)

So mental predicates are not the kind of predicates that can be given a standardized and unique definition; instead, a particular mental state, such as e.g. the belief that there is life on Mars, can be attributed to a person only in the context of *other mental* states, which in turn depend on other mental states, and so on. These features are part of the anomalous character of mental states. Since mental states cannot be given a standardized and unique definition, they do not follow a predictable chain of events – it is not the case that given certain

---

events are identical if and only if they occupy the same spatiotemporal region (Davidson, 1985). The difference between these two make no difference in regard to our discussion.

preconditions, a certain mental state will be the effect. Davidson argues that mental events do not follow any deterministic laws. However, he also claims that when there is causality, there must be a law: events related as cause and effects always fall under strict deterministic laws. This claim, in conjunction with the claim that mental states do not follow strict deterministic laws, seems to entail that mental states cannot have causal power. But Davidson argues that they do: at least some mental events have causal effect on actions. The causal power of mental states can be understood only if we appreciate the double nature of mental states. According to Davidson, mental states are mental *only by description*, but once described, this description cannot be reduced to a physical description since the mental and the physical vocabularies work in fundamentally different ways. We can conclude *a priori* that there cannot be strict deterministic laws between the physical and the mental, since such laws only hold between physical events, described in terms of physical predicates. Mental events are mental by description, but there is always a physical aspect of a mental event: the mental event is *token-identical* with some physical event. And it is in virtue of this token-identity that (some) mental events have causal power (Yalowitz, 2014).

It is not entirely clear that Pardo & Patterson would accept this solution to the mental causation problem, since it is not clear whether they accept that mental states are physical states even in a token-identity sense. But the question will now be whether *something like* Pardo & Patterson's objection (not necessarily exactly as they formulate it) to the Revision Argument works if combined with this Davidsonian view about mental causation.

5.4.2 Problem solved?

As a reminder, the underlying motivation for Pardo & Patterson's project is to show that human action cannot be thoroughly described only in terms of physical causes and effects. Daniel Dennett formulates the same worry as follows:

> The fear […] is that no naturalistic theory of the self could be given that sufficiently distinguishes it from a mere domino in a chain. We do

> not want to be mere dominoes; we want to be moral agents […] Only some of the portions of the physical universe have the property of being designed to resist their own dissolution […] only some of these have the further property of being caused to have reliable expectations of what will happen next, and hence have the capacity to control things, including themselves. And only some of these have the further capacity of significant self-improvement (through learning). And fewer still have the open-ended capacity (requiring language of self-description) for "radical self-evaluation". These portions of the world are thus loci of self-control, of talent, of decision-making. They have projects, interests, and values they create in the course of their own self-evaluations and self-definition. How much less like a domino could a portion of the physical be? (Dennett, 1984b, p. 100)

Dennett's quote nicely describes the central importance of certain mental capacities when we distinguish human behavior from mere causal interaction of e.g., a chain of dominoes. I guess that most people agree with Dennett: there is a quite obvious difference between people and domino bricks. This difference is, in itself, good evidence for why human action is not as easily described in terms of causes and effects as dominoes or bowling balls and pins. However, in this discussion, we are not primarily interested in the obvious difference between how people and domino bricks, or bowling pins, interact with the world. What we are interested in is *if we can find a ground for basic desert* in humans, complex as they are, and more specifically, if we can find an intuitively *relevant difference* between a person who is attributed with basic desert, and a person who is not.

As was pointed out above, Pardo & Patterson's approach, at least in one reading of it, seems to exclude the causal power from the nature of mental states. But it seems reasonable that mental states have to be causally effective in order to be relevant in action explanations, and also for desert-attribution. Davidson's account of mental causation was introduced in order to see if it can help save Pardo & Patterson's overall approach to mental states, by providing them with an explanation of how mental states can be causally effective even though not reducible to mental states. Davidson's approach to mental causation is connected to his view that mental states have, in a sense, a "double nature": a mental state like e.g., believing something is mental only by description, but once

described, this description cannot be reduced to a physical description since the mental and the physical vocabularies work in fundamentally different ways. Davidson claims that the different nature of these vocabularies is the reason for why mental phenomena cannot be given purely physical explanations – they cannot be combined in a way such that an explanation in physical terms can explain mental phenomena, or mental causation, for that matter. However, and importantly, Davidson does not claim that mental causation is something *other* than physical causation. He claims that insofar as mental states are involved in causal relations, this causal relation is possible because mental states are (token) identical with physical states, and physical states can be involved in causal relations.

However, this move seems to lead us back to the worry that is central in the Revision Argument. As was discussed in chapter four, the metaphysical constraint of physicalism implies that differences in functional properties must supervene on differences in physical properties. And at the physical level, it is hard to find a relevant difference between the mental state of choices, or reasons-responsiveness, or any other mental capacity that the compatibilist picks out as sufficient for basic desert, and other mental states or capacities. Even if we accept that mental concepts cannot be reduced to physical concepts, and even if we accept that different mental states have different psychological functions with regard to actions (for example, we deliberate of what to do before we make a choice, but this is not the case when we act instinctively) the physical states that are the base-properties of such mental states that (according to the view discussed) are relevant for basic desert do not seem to be relevantly different from the physical states that are the base-properties of such mental states that are not relevant for basic desert. In other words: even if we conceptualize certain mental states or capacities as relevantly different from other mental states with regard to basic desert (such as e.g., the mental capacity of reasons-responsiveness) what is going on at the physical level when someone is deliberating about what to do (i.e., she considers her reasons for action) is not relevantly different from what is going on at the physical level when someone acts on e.g., an irresistible desire, or the like. What is going on, in both cases, is that more or less complex neural

systems respond to stimuli, process it, and deliver an output in form of behavior.

It seems that in order to escape this particular problem, Pardo & Patterson need an account of mental causation according to which mental states cause action not in virtue of being token-identical with physical states, but *qua* mental states. In the next section, I turn to List & Menzies' approach to mental causation, which aims to provide such an account.

5.4.3 A difference-making account of mental causation

So far, Kim's (2000) "exclusion argument" has been an implicit restriction of the discussion of mental causation in this thesis. According to the exclusion argument, we should not have more than one complete causal explanation of the same phenomenon (except in cases of genuine overdetermination.) If mental properties are irreducible to physical properties, then it seems as that we have two complete, but different causal explanations of the same phenomenon (given that actions can be explained *both* in mental and in physical terms.) According to Kim, this double explanation creates an unstable situation requiring us to find an account of how the two purported causes are related to each other (2000, p. 65).

List & Menzies (2017) argue that the exclusion argument is unsound. If they are right, we do not need to worry about the fact that non-reductive physicalism entails that there are two complete, but independent, explanations of actions. According to List & Menzies, it is unproblematic that the very same event can be given both a complete causal explanation in mental terms, and a complete causal explanation in physical terms (although these two explanations do involve the same notion of "cause".)

List & Menzies' analysis is based on their "difference-making account of causation" according to which causation is "[...] a form of counterfactual probabilistic dependence: to be the cause of an effect is to be the difference-maker of that effect" (2017, p. 277). They spell out this thesis in counterfactual terms: *C* causes *E* if and only if two conditionals are satisfied (p. 277):

The positive conditional: If *C* were to occur, then *E* would occur.

The negative conditional: If *C* were not to occur, then *E* would not occur.

This account of causation can be contrasted against the "production account" which, for example, Kim (2000) defends. When something is a cause in the production sense, it *produces* the effect in some metaphysical sense. The cause has, as it were, causal "oomph," it is what makes the effect happen. In the difference-making account, the cause need not produce its effect in this sense. Also, as we will soon see, something can be a cause in the production sense without being a cause in the difference-making sense, since the negative conditional is not fulfilled. List & Menzies argue that the most natural way to spell out the idea of *mental causation* is to say that an agent causes an action if and only if her mental state is the difference-maker of the action (2017, p. 278). (See also List & Menzies, 2009.)

List & Menzies objection to the exclusion argument focuses of how certain concepts allow for multiple realizability, which, they argue, makes the realizer state (i.e., the physical state) unable to account for the negative conditional in the difference-making account of causation. For example, when a flask of boiling water breaks due to the pressure, they argue that the *difference-maker* is the boiling of the water, not the motion of a specific subset of water molecules. Because the boiling of the water could have been realized by a slightly different microstate, but the flask would still have broken. Had the boiling not occurred, the flask would have remained intact. Hence, the positive and negative conditionals for difference-making are satisfied when *C* is the boiling of water and *E* is the breaking of the flask. However, since it is true that the water could have boiled even if the microstructure would have been slightly different, it is not true that the flask would have remained intact if the microstructure would have been different. Hence, if *C* is the microstructure, then the negative conditional is not satisfied (List & Menzies, 2017, p. 279).

Similarly, it can be argued that, when mental states cause e.g., an action, the mental states, but not the brain states that they supervene on, satisfy both conditionals for difference-making. Since mental states can be realized by different brain states, the brain state could

have been different, and yet the same effect (e.g., the act) would have been caused. However, they argue that the base-properties – which in the boiling water case is the microstate of water molecules, and in the case with mental states is the brain states the mental states supervene upon – are *causally sufficient* for the effect (and may be causes in the production sense.) But the supervenient states – the boiling of the water, and the mental state, in the current examples –are the *difference-making* causes. This reasoning shows, according to List & Menzies, that sufficient causes at the physical level can co-exist with distinct, higher-level difference-making causes of the same effect (2017, p. 281).

5.4.4 Problem solved?

Can the difference-making account of causation provide what is needed to meet the challenge from physicalism, through its idea of how mental causation works within a non-reductive framework of physicalism? List & Menzies themselves argue in a way that suggests that they think so. They claim that what they call "the Neurosceptical Argument" fails. This argument is, in short, that since human actions and choices are completely caused by neural states and processes that are inaccessible to the agent's consciousness, human choices and actions are not free (List & Menzies, 2017, p. 280). This argument fails, they argue, since it is the mental states and processes, and not the neural states and processes, that cause human actions *in the relevant sense*, that is, the difference-making sense. This reply, in a way, parallels Pardo & Patterson's line of argument that it is a mistake to think that the causes of human actions can be found in the brain. Hence, it might be suggested that it provides what is needed to substantiate Pardo & Patterson's claim, and, thus, to refute the Revision Argument, and my argument that we have to look for relevant differences at the level of brain states. If mental states, but *not* neural states (i.e., brain states) that cause actions in the relevant sense, it seems reasonable that we do not have to look for the relevant difference between actions that are caused in a way that makes the agent responsible for them (in the basic desert-entailing sense) and actions that are not caused in that way, at the level of neural states.

I think, however, that we can, and should, resist this line of reasoning. Assume, for the sake of argument, that we accept List & Menzies' view that mental causation is best conceived of as causation in the difference-making sense. It still seems that it makes sense to ask *in virtue of what* mental states make a difference. Given physicalism, mental states cannot be involved in causal relations in a way that violates the causal closure of the physical domain. According to List & Menzies, the realizers of mental states are not the difference-makers, since mental states are multiply realizable. The fact that they are multiply realizable, in turn, depends on how mental states are identified. List & Menzies endorse functionalism about mental states, which basically means that mental concepts are such that something (let's say, some state or event) is classified as a certain mental kind, as e.g., a pain, a belief, or a desire, on the basis of its function in the overall mental system. Plausibly, having a function involves having a certain causal role which means that instances of mental states must have a certain causal effect in order to fall under a specific mental concept in the first place. Thus, before we can speak of there being mental states that can cause e.g., actions (in whatever sense) there have to be physical realizers (neural states) that can cause actions – if not in the difference-making sense so in the sense of being a sufficient cause or a production cause. Consequently, List & Menzies' view does not escape the implication that it is only in virtue of the fact that the physical realizers (neural states) cause actions (produce actions or at least as sufficient causes), that mental states cause actions.

To illustrate this point, we can make a comparison with the property of being poisonous: for something to be poisonous, it has to have certain effects on people: it will, if consumed, make them ill or die. "Being poisonous," then, is a property that is picked out via the causal effects something with the property of being poisonous has (i.e., the causal effect that someone dies or gets ill if consuming large amounts of it.) Thus, in one clear sense, it is not the fact that something has the property of being poisonous that *in itself* causes that someone dies. "Being poisonous" is, as it were, *a label* of substances with properties that (independently of the label, as it were) has the relevant causal effect. Just like e.g., "desire" (if functionalism is correct) is a label of physical states that have a certain effect in terms

of e.g., actions *independently of the label*.[48] Mental states, then, turn out to cause action in virtue of their physical realizers cause action, *precisely because* mental types are classified based on their functions. In order for a function to be in place, there must be something that makes this function happen, so to speak.

Another way to put the same point is this: List & Menzies' argue that, since mental states can be realized by many different brain states that tend to cause the same kinds of effects (actions and other things), it is to be expected that an action caused by a mental state could have been produced by many other brain states than the brain state that was actually involved. Thus, the negative conditional does not hold for particular brain states in relation to action types. But this very argument presupposes that brain states can have the relevant effects, since the brain states must have the relevant effect in order to be classified as a realizer of a certain mental type. It is only because the relevant effects can be caused by many different brain states that the brain states do not fulfil the negative conditional.

If this reasoning is sound, it seems that List & Menzie's view does not help to escape the challenges from determinism and physicalism, and, thereby, satisfy the Principle of Relevant Difference. Because also on their view, mental states cause actions in virtue of their physical realizers causing actions. Hence, when we look for the relevant difference between actions that are caused in a way that makes the agent responsible for them (in the basic desert entailing sense) and actions that are not caused in that way, it seems that we should look at the level of neural states. And then, determinism poses the same challenge as always: how can we make sense of the relevant difference between people who are morally responsible and people who are not, if everyone's actions are completely determined? If one

[48] Beebee (2017) argues in a similar manner, but she also claims also that mental states, when defined via their functional roles, end up being epiphenomena. I think she is right, but for the purposes of this discussion, it suffices to conclude that List & Menzies' difference-making account of causation cannot exclude brain-states from the discussion of mental causation, and hence, not exclude brain-states from the discussion of what the relevant difference is between mental properties that are base-properties of basic desert, and mental properties that are not.

wants to put the relevant difference elsewhere, as compatibilists usually prefer, the problem is instead the challenge from physicalism. According to this challenge, it is difficult (given the current status of neuroscience, and our current knowledge of the brain) to find a relevant difference on the level of brain states between the mental state of choices, or reasons-responsiveness, or any other mental capacity that the compatibilist picks out as base-properties of basic desert, and other mental states or capacities that are not base-properties of basic desert. Even if we accept that mental concepts cannot be translated to physical concepts, and even if we accept that different mental states have different psychological functions with regard to actions (for example, we consider our reasons for action before we make a choice, but this is not the case when we act instinctively) the brain states that realize mental states and capacities that (according to the kind of compatibilist views discussed) are the base-properties of basic desert does not seem to be relevantly different from the brain states that realize mental states and capacities that are not base-properties of basic desert. In other words: even if certain kinds of mental states and capacities on the conceptual level appear as relevantly different from other mental states with regard to basic desert (as e.g., the mental capacity of reasons-responsiveness) what is going on at the physical level when someone is deliberating about what to do (i.e., she considers her reasons for action) is not relevantly different from what is going on at the physical level when someone acts on e.g., an irresistible desire, or the like. Plausibly, different parts of the brain are involved in these different cases, but the mechanisms behind these processes are, in fundamental aspects, very similar: the way they process incoming input is completely an issue of the physical make-up of the neural networks involved, and the electro-chemical mechanisms that make these neural networks able to process information.

5.5 Summary & conclusions

In order to hold that people sometimes deserve to be punished based on how they act, human actions have to be special in some way. They have to be different from other events that do not ground desert

attributions – for example, we do not think that the bowling ball deserve blame or punishment because it happened to fall on someone's foot. In the Revision Argument it is claimed that folk psychology and folk morality take this difference to be due to that people can act freely, in contrast to e.g., a bowling ball. But, the Revision Argument goes, neuroscience gives us reason to think that people do not, in fact, act freely. A compatibilist answer to this claim is that free action is compatible with determinism since it consists in people having the right kind of capacities and abilities, and one can have these capacities and abilities even though one is completely determined to act as one does. In chapter three and four, I argued that such compatibilist theories have troubles meeting the demands of the Principle of Relevant Difference.

The Conceptual Objection discussed in this chapter, proposed by Pardo & Patterson, claims that the Revision Argument, as well as the argument I put forward in chapter 3 and 4, rest on a conceptual mistake. Even though all mental states correspond to brain-states, and all actions are caused by brain states (physicalism is correct), mental state concepts are such that mental states are not reducible to brain states. Since the relevant difference between actions that ground desert attributions and those that do not is found at the level of mental states, and such states are not reducible to brain states, it is irrelevant that there is no relevant difference at the level of brain states.

I have argued that in order to evaluate their objection, it is important to understand their theory of mental states. Pardo and Patterson themselves do not provide a positive account of mental causation, but I suggested that it is reasonable to interpret their view as an analytic, role-functionalist theory. A worry with this interpretation is that mental causation is hard to make sense of in traditional functionalist frameworks. And mental states have to be able to cause actions in order for mental state explanations to be relevant for desert attributions. Therefore, I turned to two different non-reductivist approaches to mental causation in order to scrutinize whether any of them is able to provide an account of causation combinable with Pardo & Patterson's view: Davidson's (1970)

anomalous monism, and List & Menzies' (2017) *different-making approach* to causation.

I argued that neither of these views convincingly shows why we should not look at the level of brain processes when we look for the relevant differences between natural properties (e.g., reasons-responsiveness) that are the base-property of basic desert, and natural properties that are not. Since neither of these accounts shows that mental causation can be at hand without mental states being token-identical to a physical (brain) states. And if the mental properties that are base-properties of basic desert are causally effective in virtue of being token-identical with certain brain states, it seems reasonable to involve these brain states when we are looking for the relevant difference between mental states that are base properties of basic desert, and mental states that are not.

Hence, to the extent that Pardo & Patterson's view of causation resembles any of the two views discussed here (as previously noted, they do not provide a positive account of mental causation, so I do not know to what extent they would agree with any of them) their account of responsibility meet the same requirements from the Principle of Relevant Difference, in light of the challenges from determinism and physicalism, as the compatibilist theories that were discussed in chapters three and four. If Pardo & Patterson do not want to subscribe to any particular account of mental causation, a worry is that mental events, as they are characterized in their view, end up being epiphenomena.

I conclude that the Conceptual Argument does not succeed in refuting the Revision Argument, or the arguments that I advanced against the counterfactual theories of alternative possibilities in chapter three and four.

Now, an objection to this fairly abstract reasoning that attributions of basic desert must satisfy the Principle of Relevant Difference in light of the challenges from determinism and physicalism would be that this line of reasoning does not correspond to how folk psychology and folk morality work. And both the proponents of the Revision Arguments and its critics agree that the justification of retributive punishment is, ultimately, found in folk psychology and folk morality, not in abstract philosophical reasoning that lacks

connection to how people actually think. As Al Mele writes, any adequate theory of free will [and punishment] should be "anchored by common-sense judgements" since any analysis that is completely disconnected from what the folk has to say "runs the risk of having nothing more than a philosophical fiction as its subject matter" (Mele 2001, p. 27). I fully agree with Mele in this regard. The complex relation between folk psychology, folk morality, philosophy and science is the topic of the next chapter.

6 Third objection: The limited relevance of neuroscience and philosophy for folk psychology, folk morality & the law

> Neither determinism in general nor neuroscience in particular undermines folk psychology in the ways they [Greene & Cohen] presuppose […] if moral evaluation depends on folk psychological explanations generally, and mental states exist and do causal work, then folk psychology is not illusory and provides legitimate foundation for moral evaluation. (Pardo & Patterson, 2013, p. 203)

6.1 Introduction

In the two previous chapters, two objections to the Revision Argument were discussed. According to both objections that the Revision Argument is mistaken about what role neuroscience can play for our responsibility practices. Moreover, it was argued that a compatibilist framework of free will and/or responsibility can justify legal retributive punishment. I argued that neither of these objection offers a plausible explanation of what the relevant difference is at the physical (brain) level between a person who deserves punishment and a person who does not. And, I argued, we need to be able to pick out such a relevant difference in order to justify retributive punishment, or else the conclusion of the Revision Argument stands: the retributivist element of our current legal system is unjustified.

Morse, as well as Pardo & Patterson would probably resist this conclusion, and also reject the arguments leading up to it. What I did in the previous two chapters was to provide philosophical arguments to the effect that compatibilism plausibly is not sufficient to justify retributive punishment. Morse and Pardo & Patterson might object

to this conclusion by saying that the argumentation is irrelevant with regard to the responsibility practice it is supposed to defeat.

The reason for why it is irrelevant is that the justification of the legal responsibility practice and legal retributive punishment is based on folk morality and folk psychology and, hence, in a sense, "immune" to philosophical reasoning and scientific findings. In section 3.3.1, I briefly discussed the interpretation of folk psychology in the writings of Morse and Greene & Cohen respectively, and stated that I interpret them as understanding folk psychology roughly as the ordinary person's common sense intuitions of why people behave as they do, and what considerations that can be brought into an explanation of actions. As I interpret Pardo & Patterson, they have something like this in mind when they discuss folk psychology as well.[49] This notion of folk psychology is what I will work with in this chapter, too. Neither Morse nor Pardo & Patterson explicitly employ the notion of "folk morality", but they acknowledge that people's views on moral matters are relevant for the legitimacy of the legal system. In section 6.4.2, I will distinguish between two types of folk morality, in order to show a link between folk morality and philosophy.

According to Pardo & Patterson, the folk psychological view of mental states and behavior does not presuppose anything specific about the brain. Likewise, when we ascribe moral properties to someone from a folk moral perspective, as for example when we say that someone deserves punishment because she intentionally committed a crime, this does not entail any specific belief about her brain. So even though it may be interesting and valuable to explore

[49] Pardo & Patterson write the following about folk psychology: "The expression 'folk psychology' refers to our common psychological/mental concepts and our ordinary use of words expressing these concepts. The notion of 'folk psychology' or 'folk psychological concept' is philosophical and controversial. We use the expression without endorsing the many uses to which it is put in the philosophical literature. However, the concept is a staple of the philosophical literature and, for that reason, we employ it. Nothing in our argument depends upon a rejection of the idea of folk psychology. Thus, we prescind from the direct controversy over the viability and explanatory perspicacity of this notion." (Pardo & Patterson, 2013, footnote 12, p. xviii)

the neural correlates to mental states and behavior, that we define in folk psychological terms, and evaluate from a folk moral perspective, neural correlates do not – and will not – be able to affect the folk moral view of what is required for moral responsibility. The fact that some kind of neural activity in the brain is required for mental activity (as it is defined in folk psychology) is perfectly compatible with the folk moral intuition that a certain mental capacity (defined in folk psychological terms) is required for a certain normative practice, such as responsibility attribution. Crucially, therefore, since the legal system is a folk psychological and folk moral enterprise, neuroscientific facts and philosophical arguments have very limited, if any, relevance for the law, and more specifically in this discussion, for legal retributive practices. Morse shares the view that neuroscience is largely irrelevant for folk morality and for the legal practices, since they, in turn, are fundamentally folk psychological and folk moral enterprises.

In this chapter, I will argue that folk psychology is not immune to neuroscientific information regarding mental states and their relation to brain states. In fact, such information could, at least in principle, be integrated in the folk psychological understanding of human thinking and action. Therefore, I argue, we have no conclusive reasons to think that neuroscience cannot inform, or correct, or change (current) folk psychological assumptions about human thinking and action, and that this can affect folk morality, and therefore, also legal retributive practices.

On a general level, the kind of objection to the Revision Argument to be discussed in this chapter has the following structure. (1) The justification of the retributivist element in the legal system is based on a folk psychological understanding of human behavior and folk moral approach to normative judgments, (2) Folk psychology and folk morality are resistant, or at least sufficiently resilient, to scientific information and to philosophical reasoning, (3) therefore, neuroscientific information and philosophical arguments cannot contribute to – or undermine – the justification of legal retributive punishment. I will agree with (1) but disagree with (2). Therefore, (3) does not follow.

This chapter will run as follows: in section 6.2, I will discuss different kinds of justification criteria for legal retributive punishment. In section 6.3, I will discuss the view of folk psychology defended by Morse and Pardo & Patterson and suggest an alternative view. In section 6.4, I will present two different notions of folk morality, and argue that if we appreciate this distinction, we can make sense of the fact that folk moral intuitions in some cases seem to provide contrarious responses. In section 6.5, I will summarize why and how philosophy and science are, and are not, relevant for folk psychology, folk morality and the law. I summarize the chapter in section 6.6.

6.2 Justification criteria of legal retributive punishment

In this section, different justification criteria for legal retributive punishment will be discussed. With justification criteria, I mean criteria that must be met in order for retributive punishment to be justified. If rationality is such a criterion, then it is a necessary condition that someone is rational, if we are to punish her justifiably. If libertarian free will is a justification criterion, then a necessary condition is that a person has libertarian free will in order for her to deserve punishment. When we ask what the criteria are for the justification of retributive punishment in a specific system, as for example in a legal system, this question can be interpreted in two different ways. Either, we can interpret the question as concerned with what the *legal* criteria for retributive punishment are, that is, which criteria must, *according to the law*, be met in order for retributive punishment to be justified. The legal criteria are explicitly expressed in the law as necessary conditions in order for retributive punishment to be justified. Alternatively, the question can be interpreted as concerned with the *moral* criteria of retributive punishment. According to this interpretation, the question is concerned with the moral basis of the legal practice. In a way, we can think of the latter interpretation as a question of to what extent *legal* retributive punishment is *morally* justified. These two interpretations are thus concerned with two different sets of justification criteria – criteria that

are internal to the legal practice and criteria that are external, respectively.

The idea of (external) moral justification criteria for legal punishment can (also) be interpreted in different ways. Some believe that there are objective and absolute moral facts, which then are the criteria that a legal practice (and any practice) need to satisfy in order to be morally justified, whereas others argue that moral facts are relative, or that there simply are no such facts. In this thesis I will not take a stand on any such meta-ethical position. Instead, as explained in section 2.3.1, for the purposes of the discussion in this book, I assume as a starting-point the presumption made by both proponents and critics in the debate about the Revision Argument: that the moral viewpoint that works as external justification for the legal system is the one embraced by folk morality (whatever that might be).[50]

Now, in the previous chapters, I argued that there are certain philosophical problems with regard to how compatibilism about free will and/or moral responsibility can work as the basis for basic desert attribution. But it is not *philosophical* arguments that legitimize and justify, and hence not undermine, our legal practices. The parties involved in this discussion – Morse, Pardo & Patterson and also Greene & Cohen – all agree that it is *folk morality* and *folk psychology* that do the heavy lifting in this regard.

Against this assumption, Morse and Pardo & Patterson argue that if compatibilism about free will and moral responsibility *according to*

---

[50] This should be seen as a way to delimit the discussion rather than as a substantial commitment to folk morality as the final justifier of legal systems. In other words, I do not wish to defend the view that regardless of the content of folk morality, a certain legal practice is justified as long as the practice goes in line with the content of folk morality. But in this particular discussion, I discuss the special relationship between a legal practice, folk morality and folk psychology, from the presumption that folk morality and folk psychology are necessary elements for the legitimacy of a legal system. However, I am open for the possibility that folk morality and folk psychology are not sufficient for the legitimacy or justification of a legal system, and also for the possibility that legitimacy and justification can (and perhaps should) be teared apart, so that a system may be legitimate but not justified, and vice versa. However, for the sake of simplicity, I take legitimacy and justification to co-vary with regard to the relation between folk psychology, folk morality and legal practices in the present discussion.

folk psychology and folk morality is sufficient for the moral justification of retributive punishment, then it seems that we have what is needed for a justified and legitimate legal practice. This justification holds regardless of potential philosophical worries about free will and responsibility, and regardless of scientific findings about brain states – as far as these scientific findings are not conclusive evidence to the effect that folk psychological assumptions about behavior are plainly false. For example, if neuroscience could show that people do not respond to reasons, the folk psychological assumption that people are reasons-responsive would be false. But neuroscience does not show that.[51]

In this chapter, I will challenge this view. I will not argue that neuroscience shows that folk psychological assumptions are plainly false, but I will argue that neuroscientific findings and philosophical arguments are relevant for folk psychology and folk morality *even if* the scientific findings in question do not show that folk psychological assumptions are plainly false, which seems to be what Morse as well as Pardo & Patterson require in order for neuroscience to be relevant here. Moreover, I will argue that even if compatibilism about free will and moral responsibility is sufficient for retributive punishment according to (some interpretations of) folk psychology and folk morality, philosophical arguments may change people's views about that.

The following sections will be concerned with some claims put forth by Morse and Pardo & Patterson about how the folk psychological (section 6.3) and folk moral (section 6.4) frameworks are relatively resistant to scientific explanations and philosophical

---

[51] See e.g. Morse (2007): "Laws could not guide people [...] unless people were the types of creature who could use laws as premises in their practical reasoning" (p. 205). Elsewhere, Morse writes that "folk psychology does not presuppose the truth of free will, it is perfectly consistent with the truth of compatibilism […] folk psychology presupposes and insists only that human action can at least be partially explained by mental states explanations or that it will be responsive to reasons, including incentives, under the right conditions" (Morse, 2013, p. 31).
Pardo & Patterson put it as follows: "To suppose that praise and blame require uncaused causation is to miss (or misconstrue) the normativity in human action […] neither determinism in general nor neuroscience in particular undermines folk psychology" (Pardo & Patterson, 2013, pp. 202-203).

arguments. I will argue that the view they endorse is not obviously the most plausible one, and I will provide some examples of why it is not. I will also give some examples of what I take to be a more reasonable view of folk psychology and folk morality. These examples will, in turn, affect the plausibility of the claim that folk psychological explanations can figure as justification of a legal practice even though there are scientific and philosophical arguments that show that the folk psychological explanations in question are problematic or simply false. In other words, I will provide some arguments to the effect that folk psychology and folk morality are not as resistant to scientific findings and philosophical arguments as Morse and Pardo & Patterson think they are.

6.3 Folk psychology

In this section, I will discuss two different approaches to folk psychology. In section 6.3.1, the view defended by Morse, and Pardo & Patterson will be described. According to this view, folk psychological concepts are quite resistant to the influence of neuroscientific findings. Then an alternative view will be presented. According to this alternative view, folk psychological concepts are no more resistant to neuroscientific findings than common-sense concepts in general are resistant to scientific achievements. I take the latter view to be more accurate than the former, since there is much evidence to the effect that common-sense concepts and views change in light of new information.

6.3.1 The folk psychological framework as resistant to neuroscientific explanations

According to Morse, the law's view of the person is a folk psychological:

> The law's psychology must be a folk psychological theory, a view of the person as a conscious […] creature who forms and acts on intentions that are the product of the person's other mental states, such as desires, beliefs and plans. (Morse, 2013, p. 31)

Morse argues that folk psychological concepts are necessary for the law's explanations of why people behave as they do, but also for the possibility to "adapt any morals or politics or any legal rule, or to do anything at all." In his view, a neuroscientific explanatory framework of behavior is insufficient to explain how and why people adapt their behavior to reasons, and it also leaves us with no orientation of how to navigate in normative questions:

> Normativity depends on reason, and thus the radical [revisionist] view is normatively inert. Neurons and neural networks do not have reasons; they do not have a sense of past, present, and future; and they have no aspirations. Only people do. If reasons do not matter, then we have no genuine, non-illusionary reason to adopt any morals or politics or any legal rule, or to do anything at all. Thus, this view does not entail consequentialism or a pure preventive scheme of social control. (Morse, 2013, p. 47)

Pardo and Patterson agree with Morse that neuroscientific descriptions of the brain tell us nothing about how to handle normativity, such as for example whether someone is blameworthy or not for an action. They argue that although we can find brain states that correlate to what we refer to when we say that someone has control over her actions, or we can find out what sort of brain that is required for the ability to form intentions, these findings do not, by themselves, provide an answer the question of whether someone is responsible for her actions:

> Neuro-reductionism has the effect of "flattening" [...] normative differences – differences that must be taken into account in any sufficiently adequate explanation of responsibility. (Pardo & Patterson 2013, p. 39).

Pardo & Patterson think that an important aspect of why neuroscientific explanations of actions cannot be integrated in the folk psychological explanatory framework is the very nature of the psychological concepts – even though the meanings of such concepts are not fixed, there are various criteria for applying these concepts that limit their employment. In Pardo & Patterson's view, these limitations have to do with what count as criterial evidence and what is merely inductive evidence for the phenomenon in question (2013,

p. 8). For example, they argue that the criteria for telling a lie involve behavior, not neurological states: behavior is, in their view, *criterial* evidence that someone is lying. Neuroscientific evidence could, in turn, provide us with well-grounded correlations between this type of behavior and brain states. Thus, information about brain states can constitute *inductive* evidence that someone is lying. It would, however, be a conceptual mistake to conclude that lies take place in the brain, they claim. This mistake can be illustrated if we consider a case in which a particular brain state *did* provide criteria for lies. If these criteria were met, i.e., if someone had these particular brain states, but without having the intention to lie, would we still say that the person was lying? Pardo & Patterson claim that we would not:

> What constitutes "deception" or a "lie" is a conceptual, not an empirical, question, and the criteria for the application of these concepts are behavioral, not neurological. (Pardo & Patterson, 2013, p. 101)

In sum, Pardo and Patterson argue that we must pay careful attention to how folk psychological concepts work (and do not work) in order to understand what responsibility attribution is dependent, and not dependent, on. This insight is fundamental in order to appreciate how neuroscience can be relevant for our responsibility practices. Morse does not discuss the nature of mental concepts in detail, but shares the view that responsibility attribution is connected to folk psychological explanations, and not to neuroscientific explanations, of actions.

As will be made clear in the next section, not everyone subscribes to with the view that folk psychological explanations are "immune" to neuroscientific discoveries in the sense that e.g., discoveries about the neurological underpinnings of behavior cannot be viewed as criterial evidence for mental phenomena. I will argue that even if we accept Pardo & Patterson's view that retributive practices are grounded in folk psychological concepts and explanations, this view does not exclude the possibility that the folk psychological framework can *incorporate* scientific findings of the physical basis of behavior alongside explanations in terms of mental states. In other words, nothing in their argumentation excludes the possibility for physical

facts about behavior to become part of the criterial evidence for mental phenomena.

6.3.2 The folk psychological framework as sensitive to (neuro)scientific explanations

Morse and Pardo & Patterson claim that folk psychology is necessary for our normative practices such as responsibility assessments, and that neuroscientific explanations of behavior have a very limited relevance for our folk psychological understanding of human action, and for folk moral intuitions. Neuroscientific explanations of actions are, according to Morse, "normatively inert" (2013, p. 47).

However, the views advocated by Morse, Pardo & Patterson can be disputed. One of the most renowned contemporary defenders of eliminative reductionism, Patricia Churchland, recognizes, on the one hand, the worries about the normative inertness of neuroscientific explanations of behavior, but on the other she proposes a positive perspective on what a scientific description of our behavior will mean to our social practices:

> Those who will suppose that science and humanism must be at loggerheads will greet this forecast of the future with no enthusiasm. They may tend to see the revision of folk theory and the rise of neurobiological-psychological theory as the irreparable loss of our humanity. But one can see it another way. It may be a loss, not of something necessary for our humanity, but of something merely familiar and well-worn. It may be a loss of something that, though second nature, blinkers our understanding and tethers our insight. The gain, accordingly, may be a profound increase in the understanding of ourselves, which, in the deepest sense, will contribute to, not diminish, our humanity. The loss, moreover, may include certain folk presumptions and myths that, from the point of view of fairness and decency, we come to see as inhumane. And among the desirable losses may also be a numbered certain widespread and horrible diseases of the mind-brain. (P. Churchland, 1986, pp. 482-483)

According to Churchland, a neurobiological-psychological theory (instead of "folk theory") would improve our normative practices, since it would constitute a profound deepening of our understanding of ourselves. J.Z Young provides a similar approach: "So my hope is that the application of scientific language to describe ourselves may

lead to an improvement in powers of communication and cooperation, perhaps even to a revolution in their effectiveness" (Young, 1978, cited in Churchland, 1986, p. 481). Terrence Chorvat and Kevin McCabe express their view of how neuroscientific research of behavior likely will influence our legal practices as follows:

> Research shows that human behavior is a function of a complex interaction of neural mechanisms. By understanding the neural mechanism, which we use to solve problems, we can hope to create laws and other rules that will help to foster socially optimal behavior. Such research has already given us important insights into behavior. However, future research is likely to be able to tell us how to significantly enhance compliance with the law at a minimal cost and to encourage better forms of social interaction. This research will probably completely change the way we view nearly every area of the law. (Chorvat & McCabe, 2004, pp. 127-128)

The question of how neuroscientific explanations of behavior can be of use for our normative practices, seems, hence, to elicit different intuitions in different people. Morse and Pardo & Patterson argue that neuroscientific explanations of behavior are normatively inert, whereas Churchland, Young and Chorvat & McCabe hold that neuroscientific explanations of behavior will improve our understanding of ourselves and each other, and, thereby, influence our normative intuitions about what is just and right.

If we consider these different views in light of our interest in the Revision Argument, it may, at first sight, appear that the kind of view defended by for example Churchland supports the Revision Argument, whereas the kind of view defended by Morse, Pardo & Patterson's does not. More specifically, Churchland argues that neuroscience provides us with reasons to abandon certain folk psychological explanations in favor of neuroscientific one's, and that this replacement will bring about a change in people's normative attitudes towards other people's behavior. Morse, as well as Pardo & Patterson reject this claim.

The Revision Argument claims that folk psychological explanations of actions presuppose libertarian free will, but it does not explicitly claim that folk psychological explanations, in general, are false, neither does it explicitly claim that folk psychological

explanations of actions should be *replaced* with neuroscientific explanations. An interpretation of the Revision Argument is that certain folk psychological concepts need to be adjusted in light of neuroscientific explanations of behavior, and that such adjustments will have normative implications. For example, take the folk psychological concept of fear. It is plausible that this concept has behavioral criteria for its correct application: in other words, if we are to use the folk psychological concept of fear correctly, some behavioral criteria must be met. However, according to neuroscience, certain chemical substances are always involved when we experience fear, and these substances make us behave in certain ways since they have certain bodily effects. According to Pardo & Patterson's line of reasoning, such neuroscientific findings are inductive, but not criterial, evidence of fear. But even though chemical processes in the brain have, until now, not been part of the criterial evidence of fear, why cannot they *become* part of the criterial evidence of fear?

If we scrutinize what has been common sense views, or "folk" views, in earlier times, such views have tended to change in light of new scientific achievements. For example, people believed for a long time that the earth was flat as a pancake.[52] From our everyday point of view – our "common sense" if you like – it certainly looks flat. However, at some point in history, scientific techniques made it possible for us to discover that the everyday perception of the world does not really match with the real form of the earth. It turned out that our physical size makes it impossible to directly observe that the earth actually is a sphere. This scientific finding, *even though very much in contrast to our everyday observation,* has become part of the folk understanding of the world. The idea that the earth is a sphere, is no longer considered merely a "scientific" idea, but a fact of common knowledge. It has been completely integrated and immersed in our folk view of the world, just like the idea that the earth revolves around the sun. Yet, still we refer to "sunset" and "sunrise", acknowledging our everyday experience, but simultaneously we are fully aware – on a folk level – that the sun is the center of our solar system. So, we

[52]Even today, some still believe this to be the case, though: https://www.newyorker.com/magazine/2019/04/08/its-a-flat-earth

seem to be perfectly able to revise our folk knowledge based on scientific results. The list of current folk beliefs that were once only scientific findings is long: we need the oxygen from the air; lightning is electricity; stars are far away and enormously big, in spite of how they appear.

A thorough discussion of what makes beliefs justified would be too complex to be part of this chapter, but I think it suffices to say that *if* we assume that folk psychology is a set of beliefs, these beliefs have, through history, been sensitive to new scientific information in the same way as other common sense beliefs about the world are sensitive to new information, even if the information in question is "imported" from, for example, neuroscience.

Eliminativists, like Churchland, aim to eliminate the folk psychological vocabulary and replace it with a neuroscientific vocabulary. But, if we accept the problem with multiple realizability, this replacement is a problematic approach since there may be *many* neuroscientific stories that have to replace the folk psychological "place holders". However, in my view, we need not "eliminate" nor "replace" folk psychology. I suggest that the *integration* of scientific ideas in our folk psychological framework is a more plausible option. In other words, we may allow ourselves to appreciate neuroscientific information of what is happening in the human brain when humans e.g., think rationally, *without* eliminating the folk psychological concept of rationality. So, we may enjoy the "sunset", while fully appreciating that it is the earth revolving around the sun, not the other way around.

Would it be a neuro-reductionist move to incorporate, in our everyday understanding, that e.g., sadness in humans necessarily involves low levels of certain chemical substances and high levels of other substances? Or that experiences of stress necessarily involve certain structures in the brain, such as an activated amygdala? Regardless of whether we call it reductionism or not, it seems plausible that we can learn that certain kinds of brain processes are necessarily involved when we, for example, experience fear and happiness, and that rational thinking requires certain brain structures, and so on (other processes may, for all we know, be involved when an alien experience fear, happiness, or thinks rationally). This insight must not mean that, for example, the folk psychological concept of

rationality changes referent. Rather, we could think of it as having discovered that certain brain processes are always involved when humans think rationally, and these brain processes have most likely always been involved in rational thinking. We have, in some sense, discovered something about the mental capacity of rationality, but this discovery does not mean that we have "eliminated" the folk psychological concept of rationality. It still refers to the same phenomenon, which is now *supplemented* with information.

In some contexts, such neuroscience-derived knowledge may not be relevant, but in others, it may be very helpful. For instance, if we come to know that a certain person, who suddenly has changed her personality and behaves rude and irrational, has had an accident and suffered from a head trauma, we may explain her behavior with reference to her diminished cognitive capacities. We can, of course, call such explanations for reductionism since it seems plausible to reduce the *causal relevance* of the mental states/capacities in question to the causal powers of the involved brain states (see chapter five). But, I would say, *this* sense of reductionism does not mean that we need to replace all the behavioral criteria for when someone is polite, or rational, with criteria *only* concerned with brain states.

As said, an important aspect of what neuroscience can contribute to the folk psychological understanding of behavior is that it may, in some cases, clarify *why* some people act e.g., irrationally, or have problems with controlling their behavior or act reason-responsively. For example, the phenomenon that some people act restlessly, has a hard time concentrating, and are more often than others involved in social disputes is not new. In earlier times, it may have been thought of as a sign of ill will, or perhaps lack of good rearing from the parents of a child behaving in this way. In contemporary discussions, this kind of behavior is often associated with ADHD. Symptoms similar to those of ADHD have, of course, been discussed before the neuroscientific description of this mental disorder was at hand. But to say that someone has ADHD does not only mean that she has certain functional features (for example, it is hard for her to pay attention for a longer period of time, she has difficulties with controlling her behavior in certain situations, and so on) but it may also be assumed that this behavior is related to certain impairments in brain functions, which is

information provided by neuroscience. Neuroscientific information about the brain mechanisms that are involved in the symptoms of a neuropsychiatric disorder does not wipe out our understanding of restlessness and concentration difficulties – but it adds information about why people with a certain neuropsychiatric disorder have these symptoms. Of note, much is still unknown about ADHD[53], but importantly, folk psychology is also, at least in my view, able to integrate uncertainties. Scientific information is coming to us daily by all kinds of media sources. This information, like much other information, is often accompanied by terms like "possible," "probably," etcetera. People are very much accustomed to receiving such popularized scientific information about all areas of science: history, physics, neuroscience, and perhaps most of all: medical sciences. People tend also to be aware of the facts that not all findings are eternal truth.

The fact that our folk psychological beliefs about a certain kind of behavior have been influenced and supplemented by an expanding neuroscientific understanding of the related/underlying brain functions does not exclude that we still may have certain attitudes towards someone who behaves in a way that we do not appreciate. Negative normative judgement towards people acting in ways we experience as troubling will most likely persist despite that we have knowledge of the physical basis of this behavior. However, it may also be the case that we think that certain moral responses towards people who behaves in ways we experience as "troubling" are not proper responses in light of new information. For example, to blame someone for her irrationality may seem less morally justified if we know that her lack of rationality is due to her suffering from an earlier brain trauma after a car accident. The explanation for why she behaves irrational is not that she does not pay enough attention to the circumstances, or something like it. She simply does not have the resources to think rationally, and to blame someone for her lack of such resources seems, intuitively, not to be morally right.

This reasoning has so far mostly been concerned with showing that neuroscientific facts about brains are – at least sometimes –

[53] This point certainly has to be acknowledged, see e.g., Dehue et al., 2017.

informative to our folk psychological concepts. One does not have to be a neuroscientist to consider a person's severe brain trauma when thinking of whether to attribute blame for certain behavior. Neuroscience may in some cases help us to identify people who do not deserve punishment, since we, in our current responsibility practice, do not attribute responsibility to people who lack the right mental capacities. (But, notably, I have argued that the difference between those who, according to the current practice, deserve punishment and those who do not is not of the relevant sort in order to justify this moral difference.)

Arguably, the normative relevance of neuro-information can be seen in cases when neuroscientific explanations of behavior change our intuitions about what is a morally justified reaction to that behavior. But, as Greene & Cohen point out, why are *some* differences in the neurobiology relevant for such moral intuitions, but not others? If we fully appreciate the fact that people's behavior is *completely* a consequence of their neurobiology, the apparent moral difference between people who are morally responsible, and people who are not, fades. Neuroscientific information about the causal underpinnings of the behavior apparently influences our intuitions about what is a morally justified response to certain kinds of behavior. But since neuroscientific information about the causal underpinnings of behavior can be had in *all cases of behavior,* the idea is that *if* such information influences our moral intuitions in *some* cases, it should influence our moral intuitions in *all* cases. Since arguably, the reason for why the neuro-information in question influences our moral intuitions in this way, is that if behavior is completely caused by some neurobiological processes, then it seems counterintuitive to hold that it is "up to" the person how she acts. This conclusion holds for all people, and all kinds of actions, at all times.

This above line of reasoning can be related to the point made by Pardo and Patterson: that neuroscientific explanations "flatten" the normative differences that must be taken into account in any sufficiently adequate explanation of responsibility attribution. Perhaps it is not the case that neuroscientific explanations "flatten" these normative differences, but rather, that we, when we consider

neuroscientific explanations of actions, realize that there were no grounds for such normative differences to start with.

However, according to Morse, and Pardo & Patterson, folk morality is plainly what people in general think is right or wrong, and these judgments have very little to do with what they think about brains. Therefore, facts about the brain have limited relevance for folk morality, and hence, facts about the brain have a very limited relevance for the moral justification of legal retributive punishment.

In this section, I have argued that folk psychological concepts are more sensitive to neuroscientific information than Morse and Pardo & Patterson acknowledge and that this sensitivity affects folk moral judgments. However, the claim that folk morality will change in light of such information requires a certain view of the nature of folk morality. In the next section, I will address this subject. I will suggest that the view of folk morality that is present in the writings of Morse and Pardo & Patterson is too simplistic. I will make a distinction between two conceptions of folk morality, which, I argue, makes the dismissal of the relevance of philosophy and neuroscience to the justification of current legal retributive punishment less plausible.

6.4 Folk morality

In this context, "folk morality" refers to common attitudes and responses to moral questions and dilemmas that people in general tend to have. A frequently used method to uncover folk moral intuitions is to ask people what they think about moral cases and dilemmas. In chapter three, some results from experimental philosophy were discussed, concerned with the question whether folk morality is compatibilist or incompatibilist regarding moral responsibility. Some of the results suggest that folk morality is compatibilist, whereas other results suggest that folk morality is incompatibilist.

In chapter four and five, I argued that compatibilism fails to account for an intuition that I take to be reasonable, namely, that there must be a relevant difference, or, a difference *of a certain kind* between properties that are desert entailing and properties that are not. But in what sense would a philosophical argument of that sort

be relevant to folk morality? This is the question I intend to address in this section.

To start with, the claim that there must be a relevant difference between properties that are desert entailing and properties that are not, relies on the moral intuition that (sufficiently) like cases should be treated alike. To the extent that people in general agree with me that this moral intuition is a reasonable one, then this intuition is also part of folk morality. However, the principle that there must be a difference of a certain kind, and also beliefs about which differences that are relevant and which are not, may not always be considered when people assess moral cases and dilemmas in an everyday context, or when they report their intuitions regarding moral cases and dilemmas in experimental philosophy. It is conceivable that we embrace a lot of different moral principles and fundamental beliefs that we are not necessarily aware of, or do not consider, when we engage in particular moral situations. For example, one may find oneself blaming a dog for behaving badly even if one, when keeping one's head cold, do not think that dogs are really blameworthy for their behavior – after all, they are animals that act on impulses and instincts.

6.4.1 Reactive & reflective folk morality

The line of reasoning in the section above leads to the suggestion that when analyzing folk moral responses, it makes sense to distinguish between people's "gut-reactions" in concrete moral situations, and the moral principles and beliefs that they embrace. For example, if we ask people to give moral responses to a concrete case, such as whether someone deserves punishment for something she has done, they may answer in a way that is compatibilist regarding free will, responsibility and desert. In the following section, "gut-reaction" responses to concrete cases will be discussed as *reactive* folk morality. In contrast, *reflective* folk morality encompasses moral intuitions that we have when we take a step back from our immediate reactions to particular cases and reflect on them, and the moral beliefs and principles that govern these gut-reactions. Hence, reflective folk morality includes not only moral responses to concrete cases, but also general moral principles

that people find intuitively plausible. Such a principle could be that every person is equally morally important, that it is wrong to punish someone unless she deserves it, that you should not be blamed for something you had no control over, and so on. As I think of it, reflective folk morality also encompasses intuitions about particular cases when we consider them in the light of intuitions about general moral principles and considered beliefs about the world (e.g., scientific beliefs.) It might turn out that the intuitions that we have about a particular case after such reflection are different than our spontaneous gut-reaction intuition. In other words, it is not necessarily the case that the gut-reaction responses – i.e., reactive folk morality – tracks the beliefs and intuitions embraced by reflective folk morality.

If it is true that the reactive folk morality is compatibilist regarding free will, responsibility, and desert, the philosophical arguments that I presented in chapters three and four point out a problem for reactive folk morality. I argued that given a physicalist and deterministic world-view, it seems as that there is no relevant difference between the natural properties that compatibilists put forth as base-properties of desert-attribution and retributive punishment (as for example, reasons-responsiveness), and natural properties that are not viewed as base-properties of basic desert (as for example, the ability to identify redness). In this section I will argue that these philosophical arguments are indeed based on an intuition that is already present in reflective folk morality.

Knobe & Nichols (2007) suggest that depending on the characteristics of an experimental case in, diverging moral responses will be triggered. In their study, when people are asked an abstract question about whether people can be morally responsible in a deterministic universe, this commonly triggers incompatibilist intuitions. But when the scenario instead describes a person performing a bad act (e.g., someone who is cold-heartedly killing his children), and people are asked whether he is morally responsible – still in a deterministic universe – this scenario triggers compatibilist intuitions (in this case, that he is blameworthy despite that he was determined). Their idea is that the latter kind of scenarios trigger emotions in us in a way that abstract reasoning does not. And when

people respond to examples that triggers emotional responses, they tend to give compatibilist responses, in contrast to the responses they give when they engage in abstract reasoning.

The experimental results from Knobe & Nichols show that there are measurable patterns in people's responses, but that these patterns vary depending on if people think abstractly about the scenario, or if they assess the experimental case from an emotional perspective. One way to interpret these results is that they reflect the distinction between reactive and reflective folk morality. The idea is that when we assess moral situations in an everyday context, or are presented with concrete moral experimental cases, our emotions are triggered to a higher extent compared to when we assess moral questions through engaging in an abstract mode of thinking.

But how is the distinction above relevant for the present purpose: to assess if, and if so to what extent, folk morality is sensitive to philosophical arguments and scientific facts? It is uncontroversial that folk morality is a normative enterprise, and as such, it is occupied with attitudes, judgements, and beliefs about what is right and wrong, good and bad, and so on. As mentioned earlier, one potential worry concerning the arguments in chapters three to five is that Morse and Pardo & Patterson might reply that they fail for the following reason: The legal system's understanding of behavior is folk psychological, and it is legitimized by folk morality. Consequently, as long as folk morality is consistent with a compatibilist justification of retributive punishment, then philosophical arguments of the sort that I discuss in chapters three and four – to the effect that such a compatibilist justification is problematic – are irrelevant. However, the distinction between reactive and reflective folk morality provides us with the following reply to this reasoning: it may be true that *reactive* folk morality is compatibilist in the relevant sense, but as I have argued, the Principle of Relevant Difference is part of *reflective* folk morality. Hence, folk morality seems to encompass both the intuition that a compatibilist free will can justify retributive punishment, but also the intuition that there must be a relevant difference in natural properties in order for there to be a difference in moral properties. In chapters three to five, I argued that the relevant difference condition is hard to satisfy when it comes to desert-attribution, if considering the

challenges from determinism and physicalism. Hence, philosophical reasoning makes us see that if the Principle of Relevant Difference is supported by (reflective) folk morality, then a compatibilist view of free will and/or responsibility cannot provide what is required (i.e., the relevant kind of difference) for the justification of basic desert attribution, and, therefore, folk morality does legitimize legal retributive punishment.

The distinction between reactive and reflective folk morality resembles, in a sense, of Ayer's (1962) and Russell's (2008) criticism of P.F Strawson's view of responsibility, and the justification thereof, discussed in section 4.5. Strawson argue that responsibility judgments must be understood in light of our reactive attitudes, which are " […] essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions" (Strawson, 1962b, p. 67). Ayer and Russell both object to Strawson's view that reactive attitudes do not call for justification, and argue that factual beliefs and reactive attitudes are related in a sense that Strawson fails to pay attention to. For example, if we have a reactive attitude of resentment towards someone, and this attitude involves the belief that she has done something wrong and deserves punishment, this belief is in turn connected to beliefs of what it takes to deserve something, and so on. If we realize that our notion of desert is connected to an untenable view of free will, this insight should, if we are rational creatures, affect the reactive attitude in some sense.

The disagreement between Strawson on the one hand, and Ayer and Russell on the other hand, can be understood as reflecting the difference between reactive and reflective folk morality: reactive attitudes are part of reactive folk morality, in which people's "gut-reactions" govern what seems to be proper moral responses. As Strawson points out, pure normative intuitions do not call for justification in the sense that beliefs do. Ayer's and Russell's reply to Strawson consist in that there are beliefs connected to moral intuitions, and these beliefs can be true or false. Therefore, we may be epistemically justified, or unjustified, in having these beliefs. Some beliefs may be directly connected to the moral intuition (she deserves to be punished because she did wrong, for example) but other beliefs

may not be included in, or directly inform, the normative judgments (e.g., someone deserves to be punished only if she acted freely, and in order to act freely, one must be able to act otherwise). Such beliefs that are relevant to, although not directly included in, reactive moral judgments belong to reflective folk morality. Hence, within the framework of reactive and reflective folk morality, Ayer's and Russell's criticism can be described as that they think Strawson fails to acknowledge that when we reconsider our reactive attitudes in our reflective folk morality, this reflective enterprise most plausibly affects our reactive moral judgments.

6.5 Why & how philosophy and (neuro)science are (and are not) relevant for folk psychology, folk morality and the justification of legal punishment

In the previous sections, I have elaborated on how, and why, folk psychology and folk morality may be sensitive to scientific findings as well as philosophical arguments. In this section, I will provide a summary of my arguments, and spell out the consequences this reasoning has for the Revision Argument and the objections discussed in the previous chapters.

Morse, as well as Pardo & Patterson all argue that neuroscience and philosophy have a (very) limited relevance to law. They claim that the revisionists are not only a mistaken regarding the definitions of free will and mental states, but they are also a making a more fundamental mistake by using scientific facts and philosophical reasoning in order to undermine a practice that fundamentally relies on folk psychological and folk moral intuitions. But as I have argued, both folk psychology and folk morality seem to have embedded normative standards making them sensitive to both scientific facts as well as to philosophical arguments and therefore, using science and philosophy in order to scrutinize practices that rely on folk psychology and folk morality is not a mistake at all.

In section 6.3.2, I argued that even though neuroscience may not provide us with any exact information about what kind of brain

processes that are involved in specific kinds of mental states (due to multiple realization) neuroscience can provide a rough understanding of what has to be the case in terms of brain processes in order for certain mental states to be realized. Furthermore, I mentioned some examples of when neuroscientific findings add information to why people behave as they do, and that this information in certain cases seems to alter our intuitions of what is a justified moral response to the behavior in question – not necessarily in the way that we will lose negative emotional responses towards behavior that we do not appreciate, but in the way that it may affect our intellectual reasoning about what is a *justified moral response* to such behavior. As an example, I pointed out that a person that has suffered from a brain trauma might behave in a way that evokes certain reactive attitudes in us. But if we know that such a behavior is due to a brain trauma, we may be less prone to blame her for her behavior since she may lack certain relevant mental capacities.

Another point I made was related to Pardo & Patterson's observation that neuroscience is "flattening" the normative differences that are present in folk psychology. According to Pardo & Patterson, this "flattening" effect is a reason to think that neuroscience is insufficient as a source of information with regard to our responsibility practices. I disagree. Even though I agree that neuro-scientific descriptions of behavior do not provide us with direct moral information, for example, we cannot see how certain brain processes are related to basic desert, I argue that neuroscientific descriptions of the brain can provide us with information that makes us inclined to withdraw certain moral judgments, since we, in light of the new information, no longer regard certain moral judgements as justified. More specifically, I argued in chapters three to five that we might realize that the difference between people that we judge as responsible in the basic desert sense, and people that we judge as not responsible in this sense, lack the relevant kind of difference with regard to the physical realizers of the base properties upon which

these (moral) classifications rely. Therefore, I argue, we have reason to rethink these judgments.[54]

In the section on folk morality, I argued that there are in fact two kinds of folk moral assessments: reactive and reflective. I argued further that even if *reactive* folk morality may seem to take compatibilism about free will and responsibility to be sufficient for retributive punishment, this does not entail that the same holds according to *reflective* folk morality. If the intuition that there must be a relevant difference in nature in order for a moral difference to be at hand is generally accepted, then it is a part of reflective folk morality. If so, then philosophical reasoning may help us tease out that there is a tension between the moral intuitions that are embraced by reflective folk morality, and the responses that are present in certain reactive folk moral intuitions.

However, even if I claim that philosophy and neuroscience are relevant for the justification of legal punishment (in ways that supports the Revision Argument) we should, of course, also acknowledge that there are a number of ways in which neuroscience cannot be used in the context of folk morality, folk psychology, and the law. In different places, Morse, and Pardo & Patterson have made the following points:

1. Neuroscience does not provide answers to normative questions, such as, for example, who is truly morally responsible (Pardo & Patterson, 2013, pp. 57-63; Morse, 2006, p. 400, 405).

2. Neuroscience has not shown that people do not act for reasons (Pardo & Patterson, 2013, p. 76; Morse, 2013, pp. 31-32).

[54] Note that the lack of a relevant difference with regard to basic desert does not mean the moral practice is generally unjustified. I do not think there is any problem with dividing people, actions and mental states into different moral categories, since different moral properties rely on different base-properties, and hence, we need to look for different kinds of differences depending on what moral property we are concerned with. In this particular case, I am only concerned the practice of basic desert-attribution in combination with retributive punishment.

3. Neuroscience cannot show, independently of any normative standards, that our current legal practice is unjustified (Morse, 2013, p. 46, Pardo & Patterson, 2013, pp. 186-198).

Each of these claims is fully compatible with the Revision Argument, and with the arguments that I have put forth in the previous chapters. Regarding claim (1), responsibility is, of course, not something that can be found in the brain. However, if we have a definition of what is required for responsibility, and this definition has some connection with brain states, neuroscience might be a valuable resource.

Regarding claim (2) I want to stress that I fully agree that neuroscience has not by any means shown that people do not act for reasons. The Revision Argument, and the arguments that I have put forward to its defense in this thesis, are fully compatible with the view that people are reasons-responsive and also with the claim that reasons-responsiveness has a special function in the overall cognitive system. The concern in the previous discussions has been to what extent reasons-responsiveness is a sufficient ground for retributive punishment, and I have argued that reasons-responsiveness, in combination with free will compatibilism, does not suffice as such ground, because of moral intuitions that are part of reflective folk morality. The final claim, (3), is that neuroscience cannot, independently of any normative standards, show that our current legal practice is unjustified. This claim captures the point that descriptive, factual claims about the world, such as, for example, claims about how the brain processes information, cannot in itself provide any guidance in normative matters. In order for neuroscience to be relevant in such matters, the normative practice in question must be sensitive for descriptive information of a certain kind. In this chapter, I have argued, in contrast to the third objection to the Revision Argument, that folk psychology and folk morality are sensitive to descriptive information in a way that makes neuroscience relevant for our legal responsibility practices. Hence, the conclusion of the Revision Argument stands.

6.6 Summary & conclusions

In this chapter, I have discussed what I think is a viable objection to the relevance of my arguments in chapters three to five, given the positions defended by Morse and Pardo & Patterson. This objection was divided in the following claims: (1) The justification of the retributivist element in the legal system is based on a folk psychological understanding of human behavior and a folk moral approach to normative judgment, (2) Folk psychology and folk morality are immune, or at least sufficiently resilient, to scientific information and to philosophical reasoning, (3) therefore, neuroscientific information and philosophical arguments cannot undercut the justification of legal retributive punishment. I agree with (1) but, as explained in this chapter, disagree with (2). Therefore, I have argued that (3) does not follow.

The central argument for why (2) does not hold is that I claim that folk psychology is an approach to human behavior that is much more sensitive to scientific facts than Morse and Pardo & Patterson seem to appreciate. I am not convinced by Pardo & Patterson's argument that the criteria for the correct application of mental concepts are such that they exclude information of the brain. Insofar as they are right, their argument fails to convince that facts about the brain cannot become included in the criteria for what it means to have a certain mental state, if such facts provide us with a more nuanced picture of the mental state. The problem with multiple realizability is one reason for why brain states should not be included as criteria for what it means to have a certain mental state, but as I argued, it seems reasonable to think that human brains have at least certain fundamental similarities. Such similarities can be included in our folk psychological understanding of behavior – and arguably, that kind of information is already part of folk psychology. For example, people in general (i.e., not only neuroscientists and philosophers) accept that e.g., brain trauma, neuropsychiatric disorders, etcetera, are facts about a person that contributes with explanatory relevant information about her mental states and behavior, and may change our normative attitudes towards the person.

As we all know, scientists are asked to testify in court regularly. One group that is of particular relevance to responsibility practices, is psychiatrists, particularly regarding the question of criminal responsibility and legal insanity. In their testimony, they may well refer to neuroscience, since they make use of neuroscience in their assessments. Why are scientists asked to testify in court? A plausible suggestion is that it is because folk morality demands that when we are to make morally significant decisions, we must ground them in justified beliefs, and science can add information that is relevant for us to form such justified beliefs.

As noted, the claim that neuroscience may inform us in a way that make us reconsider some of our moral judgments does not imply that we will, when having considered the neural causes of behavior, stay indifferent with regard to all kinds of behavior. We may still dislike certain kinds of behavior, and dislike people who behave in certain ways – put in Strawson's terms, we may still have reactive attitudes towards certain people and certain kinds of actions. However, these reactive attitudes, which are included in (but do not exhaust) what I call "reactive folk morality" can be scrutinized in a more reflective approach to normative questions, and the judgment that comes out of such a reflective approach is included in what I call "reflective folk morality." When arguing that our legal responsibility practices are legitimized by folk morality, we ought to take both reactive and reflective folk morality into consideration. We use our reflective moral intuitions to scrutinize our reactive moral judgments, and reflective folk morality is, or so I have argued, sensitive for scientific facts as well as philosophical arguments.

All in all, in my view, there is good evidence to conclude that the concerns I raised in the previous chapter are relevant to legal retributive punishment, and that the Revision Argument has not been refuted by the objections to it considered in this thesis.

7 Summary & concluding remarks

7.1 Introduction

In this chapter, I will provide a summary of the most important points that have been made in this thesis., and briefly discuss the consequences of my conclusions. I will close the chapter with some notes about what parts of the thesis that would benefit most from more thorough discussion, and what I would like to analyze further.

In this thesis, I have discussed some objections to what I call "The Revision Argument" which I introduced and discussed in chapter two. The Revision Argument is a version of an argument originally put forward by Joshua Greene and Jonathan Cohen in their article "For the law, Neuroscience Changes Nothing and Everything" (2004). The central message in the Revision Argument is that we lack moral justification for retributive elements in legal punishment practices. The main reason for why such justification is lacking is, according to the Revision Argument, that legal retributive punishment relies on a folk psychological understanding of actions according to which people have a libertarian free will. But, the argument goes, we have no – or insufficient – reasons to believe that people have libertarian free will, and furthermore neuroscience provides support to this belief, since neuroscience provides us with resources to explain human behavior in purely mechanistic terms. As Greene & Cohen put it, "neuroscience turns the black box of the mind into a transparent bottleneck" (Greene & Cohen, 2004, p. 1781).

According to retributivism, we are justified in punishing someone if (and only if) she *deserves* to be punished. There are different ideas of what it means to deserve something. In this thesis I have focused on basic desert, which is a notion of desert according to which someone deserves blame or punishment (or praise, or rewards) just because she has performed (a morally relevant) action, given that she has the properties required for being a person that can be blamed or punished

(or credited) for what she does. In other words, the notion of basic desert is connected to intrinsic properties of the person and her actions, and not to consequentialist or contractualist considerations. Basic desert must, as I have argued, be in place in order for retributive punishment to be justified. To be morally responsible in the basic desert sense is to be such a person that can deserve, in the basic desert sense, blame and punishment when she commits wrongful actions.

According to the Revision Argument, legal practices are legitimized by folk psychology and folk morality. Furthermore, our current legal retributive practice is justified with reference to a folk moral view of responsibility that requires that people have a libertarian free will in order to deserve punishment. This claim is highly contested, and in this thesis, I have discussed three objections to it, to the effect that the Revision Argument is mistaken about that folk psychology and folk morality require libertarian free will for retributive punishment to be justified. All these objections accept physicalism and determinism as constraints in the discussion of what makes retributive punishment justified. In other words, they aim to defend the view that retributivism can be justified within a compatibilist framework of free will and/or moral responsibility.

In order to justify retributive punishment within a compatibilist framework, constrained by determinism and physicalism, it must be possible, I have argued, to specify some intrinsic property belonging to the desert subject that is sufficient for attributing the moral property of basic desert to her. Furthermore, I argued that the property that is picked out as sufficient for basic desert must satisfy what I call "the Principle of Relevant Difference" (this goes for any theory, both compatibilist and incompatibilist ones.) This principle is based on the moral intuition that two states of affairs cannot have different moral properties (or, in other words, different ethical character) without being relevantly different with regard to natural properties, since moral differences in a physicalist framework supervene on natural differences. Further, it is also intuitively plausible that not *any* natural difference will do as the supervenience-base of a moral difference: in order for two states of affairs to differ with regard to moral properties, there must be a natural difference between them *of a relevant kind*. This is the content of the Principle of

Relevant Difference (introduced in chapter three). The moral intuition that there must be a natural difference of the relevant kind between two states of affairs in order for there being an ethical difference between them, is also related to the moral intuition that I, following Kim (1984), call "the consistency requirement" which is the requirement that like cases should be treated alike. In this thesis, I have specified this to be the requirement that *sufficiently* like cases should be treated alike – as was pointed out in relation to the relevant difference condition, not all natural differences between two cases are morally relevant in the sense that they plausibly make a difference to the ethical character. This means that the consistency requirement does not only cover two states of affairs that are alike in *all* aspects – it also covers cases that are *sufficiently similar*.

Since the involved parties in this discussion accept physicalism and determinism, these metaphysical doctrines restrict what can be picked out as a relevant difference with regard to basic desert. In chapter three, I introduced and discussed these restrictions as "the challenge from determinism" and "the challenge from physicalism", respectively. In chapter four and five, I discussed whether the objections to the Revision Argument put forward by Morse and Pardo & Patterson can satisfy the Principle of Relevant Difference and meet the challenges from determinism and physicalism. In the two sections to come, I will summarize my arguments to the effect that the objections to the Revision Argument that I have discussed in this thesis fail to meet these challenges and satisfy the Principle of Relevant Difference.

7.2 The challenge from determinism

The challenge from determinism is perhaps the most commonly discussed challenge to compatibilist theories of free will and moral responsibility. The central worry of this challenge is that if all events, both mental and physical, are equally determined, it is hard to see how it would make sense that some of them are such that the agent having them could be held morally responsible for them in the basic desert sense. In chapter four, Van Inwagen's (1981) consequence argument was introduced as a version of this challenge. The Consequence

Argument more specifically states that if we are not responsible for the facts that lead up to what we do (i.e., we are not responsible for how the world has been, from the distant past up to the time of our action) we cannot be responsible for our actions or the consequences of our actions, either. Some proponents of PAP – The Principle of Alternate Possibilities – argue that determinism blocks the, for free will and/or moral responsibility, necessary requirement of alternative possibilities, and that therefore no one can be morally responsible in the basic desert sense.

As was discussed in chapter four, Frankfurt (1969) as well as Fischer (2002, 2012) defend a kind of compatibilism that I called "actual sequence compatibilism" according to which determinism, in fact, is not a threat to moral responsibility, since in their view, we should reject PAP. They argue that what really matters for moral responsibility is the kind of psychological process that leads to the action – in Fischer's vocabulary "the actual sequence" that is involved when the action comes about – not whether we have a genuine possibility to do otherwise than we actually do. For example, if the psychological mechanism that leads to an action is reasons-responsive, then one might be morally responsible for the action, which is not the case if the mechanism was not reasons-responsive.

Reasons-responsiveness may, at a first glance, indeed seem to be a relevant difference with regard to basic desert. However, when we scrutinize the "actual sequences" that led to action in two persons who both have committed a criminal action, it is hard to see the relevant difference between them with regard to basic desert attribution, even if, at the time of the action, one of them was reasons-responsive and the other was not. The sequences in both persons were very similar in the respect that they, under the circumstances at hand, were determined to lead to a criminal action. (Admittedly, this conclusion is most likely not accepted by the compatibilist. I return to this issue shortly.)

I also discussed some compatibilist approaches that I called "counterfactual theories of alternate possibilities" according to which PAP is a necessary requirement for moral responsibility, and that e.g., reasons-responsiveness constitutes a relevant difference with regard to basic desert since it meets the requirement from PAP. According

to these theories, what it means to have an alternate possibility is that one, at the time of an action, had the capacity to do something else *in a nearby possible world*. In response to such theories, I argued that a dispositional property cannot do any causal work in relation to the actual action. And when we look at the actual physical processes that lead to an action in *this* world, the criticism that I put forward to Frankfurt's and Fischer's accounts, applies to these theories as well.

In chapter five, I discussed another objection to the Revision Argument according to which the assumption that determinism threatens moral responsibility presupposes reductive physicalism about mental states. According to the objection, reductive physicalism is an untenable view of mental states, since it is based upon a confused understanding of mental concepts and if these confusions are straightened out, it will become clear that the moral judgments connected to mental state explanations of actions are not challenged by determinism. I argued that this objection has troubles escaping the challenge from physicalism, and I will summarize why in the following section.

7.3 The challenge from physicalism

If determinism is correct, then the difference between actions that the agent is responsible for (in the basic desert sense) and other actions (and non-actions), cannot be that the former actions (or the choices that precede them) are *not* caused, but rather that the different actions are caused by different mental processes and states. For example, according to one popular suggestion the relevant characteristic property of a responsible agent (that is, the base property of basic desert) is that her actions are based on mental processes that involve reasons-responsiveness or rationality. So, the relevant difference has to do with how the action is mentally caused. Further, if we accept physicalism, then mental causation must be found in physical processes since, according to the causal closure of the physical, all physical effects must have physical causes. Therefore, the difference between people that we think of as responsible (in the relevant sense) and others, has to do with the fact that their actions are caused by different kinds of brain-processes (or, in other words, brain processes

with different roles in the overall cognitive system). But when we look for a difference at the physical level of brain processes, it is hard to find one which is intuitively relevant with respect to basic desert. Because all brain processes work in accordance with the same input-output principles: they are sensitive to input, they are processing input in a certain way depending on the preconditions at hand, and deliver output. All outputs are products of incoming stimuli in combination with the preconditions in the system that receives and processes the input.

Without a relevant difference, choosing one brain process as the base for basic desert, over another that seems sufficiently similar, would violate the consistency requirement. This is the challenge from physicalism. At this moment in time, there is no scientific or philosophical theory available that explains what constitutes the relevant difference regarding blameworthiness at the brain level.

However, the claim that the Principle of Relevant Difference must be satisfied at the level of brain states is contested. One objection to this line of reasoning is that, even if physicalism is correct, and even if responsibility attribution, and basic desert, is intimately connected to mental state explanations of actions, such explanations cannot be reduced to brain state explanations. Therefore, the relevant difference between someone who is morally responsible and someone who is not must be explained in terms of mental concepts, and mental concepts are irreducible to physical concepts.

In chapter five, I discussed Pardo & Patterson's (2013) argument to this effect. If we are to assess their argument, I argued that we must understand their theory of mental states, but they do not elaborate their view on this matter in much detail. In order examine their argument, I suggested that their view had similarities with Donald Davidson's "anomalous monism" according to which mental explanations are irreducible to physical explanations. This irreducibility is due to multiple realizability of mental states, which gives mental states their anomalous character. However, Davidson accepts that the causal efficacy of mental states must be found at the level of brain states and, which means that the problem returns: at that level it is hard to satisfy the relevant difference condition. Hence, Pardo & Patterson's theory require, in order to not be vulnerable for

the same criticism as was directed towards the compatibilist theories discussed in chapter four, an account of causation that does not place the causal efficacy at the level of brain states. One such theory is defended by List & Menzies (2017), who argue that the most plausible account of causation is a *difference-making account.* According to such an account, an event is the cause of an effect if and only if is satisfies both a positive and negative conditional. *C* causes *E* if and only if: if *C* were to occur, then *E* would occur *and* If *C* were not to occur, then *E* would not occur (List & Menzies, 2017, p. 277). According to List & Menzies, even if both conditionals are satisfied by a mental event causing an action (*E*), only the positive is satisfied by the brain event that realizes the mental event. Since mental events are multiply realizable, the actual physical realizer event could have been absent, and *E* would still have occurred since the mental event could have been realized by some other physical event.

I argued that List & Menzie's account of mental causation does not escape the challenge from physicalism. Rather, I claim, List & Menzie's account actually *presupposes* that it is brain states that are causally efficacious in mental causation. They claim that the reason for why mental states, but not brain states, fulfill the negative conditional is that mental states are multiple realizable, that is, a mental type can be realized by many different brain states. That mental states are multiply realizable is, in my view, an unproblematic assumption, but the fact that they are multiply realizable, in turn, depends on how mental states are identified. List & Menzies endorse functionalism about mental states, which basically means that mental concepts are such that something (let's say, some state or event) is classified as a certain mental kind, as e.g., a pain, a belief, or a desire, on the basis of its function in the overall mental system. Plausibly, having a function involves having a certain causal role which, in turn, means that tokens of mental states must have a certain causal effect in order to fall under a specific mental concept, or to be identified as a specific mental type, in the first place. Thus, before we can speak of mental states that can cause e.g., actions (in whatever sense) there has to be physical realizers (neural states) that can cause actions – if not in the difference-making sense so at least in the sense of being a sufficient cause or a production cause. Consequently, List & Menzies'

view, I argued, does not escape the implication that it is only in virtue of the fact that the physical realizers (neural states) cause actions (produce actions or at least as sufficient causes), that mental states cause actions. This means that the causal efficacy, also in their view, is located at the level of brain states, and that we, therefore, must seek for relevant differences between cases where we attribute moral responsibility and cases where we do not, at the level of brain states. Consequently, the challenge from physicalism is not tackled.

In sum, non-reductive physicalism of the kind Pardo & Patterson defend falls prey to the same criticism that was directed towards the compatibilist theories in chapter four: in order to satisfy the Principle of Relevant Difference, we need to find a relevant difference between the brain states that play the causal role in events for which we attribute moral responsibility, and brain states that play the causal role in events for which we do not. But as I have argued, it is difficult to find such a relevant difference.

Some people may find it convincing that "we just do not know at this moment in time, but we may find the brain-level difference between blameworthy and non-blameworthy actions in some point in time in the future." This may very well be the case. However, what I have argued for in this thesis is that at this point in time, we have not discovered such a difference. And in my view, if we want to justify our current retributive punishment practices – which involves intentionally, and often seriously, harming another person *only because he or she allegedly deserves it* – today, we have to be able to justify this practice referring to what we *currently* know, not on what we may (or may not) know in the future. That is the assumption that motivates my analysis about the Relevant Difference. I realize that some may find it a heavy burden of proof that I place on retributivist practices, but intentionally harming other people merely because they "deserve it", requires a strong and clear basis in order to be justified.

7.4 Folk psychology, folk morality and the justification of retributive punishment

A possible objection to my arguments in chapters three to five is that legal retributive practices are not legitimized by philosophy and

science: a legal practice is, rather, legitimized by folk psychology and folk morality. In chapter six, I discussed this objection in form of the following claims: (1) The justification of the retributivist element in the legal system is based on a folk psychological understanding of human behavior and folk moral approach to normative judgment (2) Folk psychology and folk morality are immune, or at least sufficiently resilient, to scientific information and to philosophical reasoning, (3) therefore, neuroscientific information and philosophical arguments cannot undercut the justification of legal retributive punishment. If this is correct, one might claim that both neuroscience, as well as my *philosophical* arguments in chapter three to five to the effect that retributivism cannot be based on compatibilism, fail to show anything about the actual justification of retributivism in the legal system. I argued against this conclusion. I agree with (1) but disagree with (2). Therefore, I argued that (3) does not follow.

The central argument for why (2) does not hold is that folk psychology is an approach to human behavior that is much more sensitive to scientific facts than Morse and Pardo & Patterson seem to acknowledge. For example, information about the neurobiological underpinnings of behavior may influence, and change, folk psychology and folk morality, even though neuroscience does not show that the folk psychological beliefs that we had before we knew anything about these underpinnings were plainly false. I claim that information about what is happening in the brain when having certain kinds of mental experiences may become integrated in the criteria for what it means to have a certain mental state, if this provides us with a more nuanced picture of the mental state. The problem with multiple realizability is one reason for why brain states should not be included as criteria for what it means to have a certain mental state, but as I argued, it seems reasonable to think that human brains have at least certain fundamental similarities. Such similarities can be included in our folk psychological understanding of behavior – and arguably, that kind of information is already part of folk psychology: people in general (i.e., not only neuroscientists and philosophers) accept that e.g., brain trauma, neuropsychiatric disorders, etcetera, are facts about a person that contribute with explanatory relevant information about her mental states and behavior.

With regard to folk morality, I distinguish between "reactive" and "reflective" folk morality. Reactive folk morality consists of "gut-reaction" responses to moral questions and dilemmas, whereas reflective folk morality encompasses moral intuitions that we hold when we take a step back from our immediate reactions to particular cases. As I think of it, reflective folk morality is in play when we consider particular moral questions and dilemmas in the light of intuitions about general moral principles (such as, for example, the principle "treat like cases alike"); in the light of considered beliefs about the world (e.g., scientifically informed beliefs, e.g., that certain illnesses give raise to deviant behavior); and in the light of what we take as relevant distinctions (e.g., distinctions between different moral categories such as wrongness and blameworthiness, or subjective and objective wrongness.) It might turn out that the intuitions that we have about a particular case after such reflection is different than our spontaneous gut-reaction intuition, i.e., it is not necessarily the case that the gut-reaction responses – i.e., reactive folk morality – tracks the beliefs and intuitions embraced by reflective folk morality.

If it turns out that reactive folk morality is compatibilist regarding free will, responsibility and desert, the philosophical arguments put forth in chapter three to five point out a problem for this part of reactive folk morality. Moreover, I argued that this critique is not a philosophical argument that is completely detached from folk morality. Rather, it is supposed to be an argument that stems from another folk moral intuition – an intuition that belongs to reflective folk morality. The fact that different scientific experts are, on a regular basis, asked to testify in court points, or so I argued, to the fact that according to folk morality, we should ground morally significant decisions on justified beliefs. And science can add information that is relevant for us when forming such justified beliefs.

Given this view of folk psychology and folk morality, there is, in my view, good evidence to conclude that the concerns I raised in the previous chapters are relevant to legal retributive punishment, and that the Revision Argument has not been refuted by the objections to it considered in this thesis.

7.5 Implications for the legal system

Given that the compatibilist theories that I have discussed in this thesis fail to provide a base property of basic desert that satisfy the Principle of Relevant Difference, the Revision Argument is not refuted and the retributive element in the legal system lacks justification. What has been picked out as the base-properties of basic desert (I have mostly focused on reasons-responsiveness) does not meet the requirements of the Principle of Relevant Difference in light of the challenges from determinism and physicalism.[55]

As described in section 2.3.2, the notion of desert that has been relevant for the purposes of this book is "basic desert" (following e.g., Pereboom, 2014 and Smilansky, 2000) and the notion of moral responsibility that has been in focus has been "retributive desert moral responsibility" (following Caruso & Morris, 2017.) It has been a presupposition throughout this book that it is this kind of responsibility, and this kind of desert, that has to be in place in order for retributive punishment to be justified. However, it seems plausible that there are other notions of moral responsibility that are not essentially connected to desert. As was pointed out in section 2.5, the view that moral responsibility and justified punishment can be had without accepting that people deserve anything in the sense of basic desert has been embraced by a number of philosophers, as for example J.J.C Smart (1961), Daniel Dennett (1984b), Saul Smilansky (2000) and Derk Pereboom (2014). David Hume can also be interpreted as defending this view (Russell, 1990, p. 560). In such a responsibility practice, what makes moral responsibility attribution justified, as well as the subsequent response (blame, praise, punishment, reward, and so on) is that this practice has consequences we consider valuable. In such a responsibility practice, basic desert has no place, but there may be room for what has been called *derived*

[55] It might, of course, be the case that there are compatibilist theories that I am unaware about, that can satisfy the Principle of Relevant Difference, as well as it might be features of the brain that we don't know about yet, but that satisfies the relevant different condition on the level of brain states.

desert (see e.g., Pereboom 2001, 2014.)[56] Within such a practice, punishment may be justified in a similar way as in a practice in which punishment is justified with reference to basic desert– it is justified to punish someone only if she deserves to be punished – but the desert that has to be in place in order for punishment to be justified is derived desert, not basic desert. The fundamental difference between these two notions of desert is that the base property of derived desert is not an intrinsic property of the person – that is, that she has a property that is sufficient for her to deserve punishment just on the basis of what she has done – but rather a relational property: the property of being a person (or a kind of person) that it has desirable consequences to punish.

For the purpose of this thesis, it is important to keep these two notions of desert apart. Pereboom (2014) points out that some philosophers identify themselves as compatibilists because they think that some non-basic notion of desert and responsibility is compatible with determinism. For example, Dennett says that his compatibilist notion of free will "can play all of the valuable roles free will has traditionally been invoked to play" (Dennett 2003, p. 225, cited in Pereboom, 2014, p.3), and Jackson expresses a similar idea, stating that compatibilist arguments do not show that "[…] free action as understood by the folk is compatible with determinism, but that free action on a conception near enough to the folk's […] does the theoretical job we folk give the concept of free action in adjudicating questions of moral responsibility and punishment, and in governing our attitudes to the actions of those around us, is compatible with determinism" (Jackson, 1998: pp. 44-45, cited in Pereboom, 2014, pp. 2-3). But, as Pereboom notes, if compatibilism is defined so that also non-basic notions of desert and moral responsibility are compatibilist, "[…] virtually everyone in the debate stands to be a compatibilist" (Pereboom, 2014, p. 2). A reason for why this is problematic is that it is hard to identify the disagreement between those who argue that basic desert is possible, and those who argue that it is not, if we do not keep these notions of desert apart, and discuss both views under

[56] Feinberg (1970) also talks about "basic" and "derived" desert, but the distinction that I have in mind here is different from his.

the label of compatibilism. And there is no doubt that there is a disagreement regarding moral responsibility, what it is and what it requires. And, perhaps more importantly, it is of moral importance to be clear about what kind of desert we have in mind when punishing people in the legal system, if one notion of desert is likely to be unjustified. As Jackson and Dennett note, basic desert and derived desert may provide similar outputs from one point of view, but they have fundamental moral differences. These similarities and differences can be identified when distinguishing between what Moore calls "internal" and the "external" questions (Pereboom, 2014, p. 49). Rawls defends a similar distinction, arguing that it is important to distinguish between *justifying a practice*, and *justifying a certain action falling under it* (Rawls, 1955, p. 3). As Rawls notes, this distinction has "frequently been made" and is central to works of Hume, Austin, Mill, and others (Rawls, 1955, footnote 2, p. 3).

If we return to the discussions of this thesis, I have argued that a compatibilist justification of legal retributive punishment lacks justification. The justification that I have in mind is what Moore calls "external" justification, or what Rawls talks about as "justifying a practice." However, it does not necessarily have any fundamental implications for how the "internal" justification looks like, or in Rawls' words, how we justify the specific act of punishment within the legal punishment practice.

If we think that the practice of punishing people with reference to their desert has more desirable consequences than it would be to give up this practice, we can (justifiably) go on doing this also if we accept that no one has the properties required for basic desert: we can replace the notion of basic desert with the notion of derived desert. The relevant difference between the one who deserves to be punished (in the derived sense) and the one who does not must then be that the one who deserves to be punished has a property such that it makes her a person that, if she is punished for her misdeeds, this will lead to certain desirable consequences. Given that the law is an action-guiding enterprise, reasons-responsiveness might play an important role when we distinguish between those who deserve (in a derived sense) punishment, and those who do not. Because people who are reasons-responsive may be more prone to change their behavior in

the desired way when exposed to certain kinds of interventions (such as punishment), compared to people who are not reasons-responsive. If so, reasons-responsiveness may be sufficient for desert, not because it is relevantly different from other capacities with regard to its nature, but because it is required in order for legal punishment to have the intended consequences.

The debate about the implications of neuroscience for criminal law is likely to continue. I hope to have contributed to this discussion by showing that the Revision Argument is probably more relevant than sometimes acknowledged. In addition, I have sketched a view of folk psychology and folk morality that allows for these enterprises to be more open for scientific findings and philosophical arguments than is appreciated by Morse as well as Pardo & Patterson. This openness has, in turn, implications for the argument that the law is a folk psychological and folk moral enterprise, in the sense that although this may very well be the case, these frameworks allow for, and in some cases even require, that scientific findings and philosophical reasoning are employed in order to reinforce and legitimize our everyday moral practices.

7.6 Suggestions for future research

I have argued that legal retributive punishment based on basic desert cannot be justified in a compatibilist framework. In my arguments, I have presupposed determinism and physicalism. Of these two, I take physicalism to be the least controversial. However, I think that a more detailed definition of physicalism may contribute to the clarity of my argument, and to the general thesis that moral facts supervene on natural facts. In section 1.5.2, I provided a (very) brief overview of some libertarian accounts of free will. I have not related any of the discussions in this thesis to such a libertarian account, but I think it would be fruitful to analyze how different libertarian accounts of free will in relation to moral responsibility and basic desert can handle the Principle of Relevant Difference in light of the challenge from physicalism.

The arguments put forward in this thesis hinges to a large extent on the notion of mental causation. Mental causation is a well-

discussed philosophical issue and there are many aspects of this discussion that have not been addressed in this thesis. A detailed account of mental causation requires metaphysical footwork – causation in itself is a much-disputed phenomena, and when we discuss *mental* causation, we add the much-disputed nature of the mental – and I have not had the space for elaborating on these discussions in this thesis. However, as was illustrated in the discussions concerning mental causation and functionalism, it is important to know what one is looking for in one's theory of mental causation. For example, is it, as in List & Menzie's theory, intended to describe the way we think of causation given the multiple realizability of mental concepts, or are we interested in the substances that do the causal work, regardless of how the concepts work? In future work, I think that my arguments in this thesis could be further developed within a more carefully defined view of mental causation.

References

Aristotle. (2017). *De Anima* (C. D. C. Reeve, Trans.). Indianapolis, CA: Hackett Publishing Company, Inc.

Armstrong, D. M. (1968). *A Materialist Theory of the Mind* (2 ed.). London: Routledge.

Armstrong, D. M. (1981). *The Nature of Mind*. Brisbane, AU: University of Queensland Press.

Ayer, A. J. (1954). Freedom and Necessity. In R. Shafer-Landau (Ed.), *Ethical Theory (2013)*. Chichester, West Sussex; Malden, MA: Wiley-Blackwell.

Ayer, A. J. (1962). Free Will and Rationality. In M. McKenna & P. Russell (Eds.), *Free Will and Reactive Attitudes (2008)*: Ashgate Publishing Limited.

Bagaric, M., & Amarasekara, K. (2000). The Errors of Retributivism. *Melbourne University Law Review, 24*(124).

Beebee, H. (2017). Epiphenomenalism for Functionalists. In H. Beebee, C. Hitchcock, & H. Price (Eds.), Making a Difference: Essays on the Philosophy of Causation. Oxford Scholarship Online: Oxford University Press.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell

Björnsson, G., & Persson, K. (2011). The Explanatory Component of Moral Responsibility. *Noûs, 46*(2), 1-29.

Björnsson, G., & Persson, K. (2012). A unified empirical account of responsibility judgements. *Philosophy and Phenomenological Research, 87*, 611-639.

Block, N. (1994). Functionalism. In S. Guttenplan (Ed.), *A Companion to the Philosophy of MInd*. Oxford: Blackwell Publishers Ltd.

Block, N. (2007). Max Black's Objection to the Mind-Body Theory. In T. Alter & S. Walter (Eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Block, N., & Fodor, J. (1972). What Psychological States Are Not. *Philosophical Review, 81*, 159-181.

The Brain Initiative. (2019). Retrieved from https://www.braininitiative.org/mission/ April 29, 2019

Broad, C. D. (1925). *The Mind and its Place in Nature*. London: Routledge & Kegan.

Burge, T. (1993). Mind-Body Causation and Eplanatory Practice. In J. Heil & A. Mele (Eds.), *Mental Causation*. Oxford: Clarendon Press.

Catley, P. (2016). The Future of Neurolaw. *European Journal of Current Legal Issues, 22*(2).

Catley, P., & Claydon, L. (2015). The use of neuroscientific evidence in the courtroom by those accused of criminal offenses in England and Wales *Journal of Law and the Biosciences, 2*(3), 510-549. doi:https://doi.org/10.1093/jlb/lsv025

Chalmers, D. J. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chandler, J. A. (2015). The use of neuroscientific evidence in Canadian criminal proceedings *Journal of Law and the Biosciences, 2*(3), 550-579. doi:https://doi.org/10.1093/jlb/lsv026

Chisholm, R. (2015). Human Freedom and the Self. In J. Dancy & C. Sandis (Eds.), *Philosophy of Action. An Anthology* (pp. 347-352). Oxford: Wiley Blackwell.

Chorvat, T., & McCabe, K. (2004). The brain and the law. In S. Zeki & O. Goodenough (Eds.), *Law & the Brain* New York: Oxford University Press.

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

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of philosophy, 78*(2).

Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

Clarke, R., & Capes, J. (2017). Incompatibilist (Nondeterministic) Theories of Free Will. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Speing 2017 ed.): Metaphysics Research Lab, Stanford University.

Cottingham, J. (1979). Varieties of Retribution. *The Philosophical Quarterly, 29*(116), 238-246.

Cupit, G. (1996). Desert and Responsibility. *Canadian Journal of Philosophy*(26), 83-100.

Davidson, D. (1963). Actions, Reasons and Causes. *Journal of philosophy, 60*(23), 685-700.

Davidson, D. (1969). Actions, Reasons and Causes. In D. Davidson (Ed.), *Essays on Actions and Events*. New York: Oxford University Press.

Davidson, D. (1970). Mental Events. In D. J. Chalmers (Ed.), *Philosophy of Mind, classical and contemporary readings*. New York: Oxford University Press

Davidson, D. (1985). Reply to Quine on events. In E. Lepore & B. McLaughlin (Eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. New York: Blackwell.

de Kogel, C. H., & Westgeest, E. J. M. C. (2015). Neuroscientific and behavioral genetic information in criminal cases in the Netherlands *Journal of Law and the Biosciences, 2*(3), 580-605. doi:https://doi.org/10.1093/jlb/lsv024

Dehue, T., Bijl, D., de Winter, M., Scheepers, F., Vanheule, S., van Os, J., . . . Verhoeff, B. (2017). Subcortical brain volume differences in participants with attention deficit disorder in children and adults. . *Lancet Psychiatry, 4*(6), 438-439.

Dennett, D. (1984b). *Elbow Room: The Varieties of Free Will Worth Wanting*. . Cambridge: MIT Press.

Dennett, D. (2003). *Freedom Evolves*. London: Penguin Books.

Duff, A. R. (1990). Justice, Mercy, and Forgiveness. *Criminal Justice Ethics, 9*(2), 51-63.

Duff, A. R., & Hoskins, Z. (2018). Legal Punishment. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Fall 2018 ed.): Metaphysics research Lab, Stanford University.

Earman, J. (1986). *A primer on determinism*. Dordrecht: D. Reidel Publishing Company.

Farah, M. J., Hutchinson, B. J., Phelps, E. A., & Wagner, A. D. (2014). Functional MRI-based lie detection: scientific and societal challenges. *Nature Reviews Neuroscience, 15*(2), 123–131.

Farahany, N. A. (2015). Neuroscience and behavioral genetics in US criminal law: an empirical analysis *Journal of Law and the Biosciences, 2*(3), 485-509. doi: https://doi.org/10.1093/jlb/lsv059

Feigl, H. (1958). The 'Mental' and the 'Physical'. In H. Feigl, G. Maxwell, & M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science: Concepts, Theories and the Mind-Body Problem*. Minneapolis: University of Minnesota Press.

Feinberg, J. (1970). Justice and Personal Desert. In J. Feinberg (Ed.), *Doing and Deserving: Essays in the Theory of Responsibility*: Princeton University Press.

Feltz, A., & Cokely, E. T. (2009). Do judgements about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 18*, 342-350.

Feyerabend, P. K. (1963). Comment: Mental Events and the Brain. *Journal of philosophy, 60*(11), 295-296.

Fischer, J. M. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.

Fischer, J. M. (2002). Frankfurt-style Examples, Responsibility and Semi-compatibilism. In R. Kane (Ed.), *Free Will*: Blackwell Publishers.

Fischer, J. M. (2012). Semicompatibilism and Its Rivals. *The Journal of Ethics, 16*(2), 117-143.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, UK: Cambridge University Press.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. (2002). Special Sciences (or: The Disunity of Sciences as a working Hypothesis). In D. J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings* (pp. 126-135). New York: Oxford University Press.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy, 66*(23), 829-839.

Gazzaniga, M. (2006). Facts, fictions and the future of neuroetichs. In J. Illes (Ed.), *Neuroetichs: Defining issues in theory, practice and policy*. (pp. 141-148). Oxford, UK: Oxford University Press.

Ginet, C. (1966). Might We Have No Choice? In K. Lehrer (Ed.), *Freedom and Determinism*. New York: Random House.

Ginet, C. (1989). Reasons Explanations of Action: An Incompatibilist Account. *Philosophical Perspectives, 3*, 17-46

Golding, M. P. (1975). *Philosophy of Law*: Englewood Cliffs.

Greely, H. T. (2011). Reading minds with neuroscience – Possibilities for the law. *Cortex, 47*(10), 1254-1255.

Hall, K. L. (2004). *The Oxford Companion to American Law*. In.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Harris, S. (2012). *Free Will*. New York: Free Press.

Hart, H. L. A. (1968). IX: Postscript: Responsibility and Retribution. In *Punishment and Responsibility* (pp. 210-237). Oxford: Clarendon Press.

Hellman, G., & Thompson, F. (1975). Physicalism: Ontology, Determination and Reduction. *Journal of philosophy, 72*(17), 551-564.

Human Brain Project. (2017). Retrieved from https://www.humanbrainproject.eu/en/about/overview/ April 27, 2019

Hume, D. (2019). An Enquiry Concerning Human Understanding. In R. Ariew & E. Watkins (Eds.), *Modern Philosophy. An Anthology of Primary Sources* (Third ed., pp. 579-646). Indianapolis, Cambridge: Hackett Publishing Company.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford Oxford University Press

Kane, R. (1985). *Free Will and Values*. Albany: State University of New York Press.

Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. (2000). The Dual Regress of Free Will and the Role of Alternative Possibilities. *Philosophical Perspectives, 14*, 57-79.

Kane, R. (2002). The Contours of contemporary Free Will Debates. In R. Kane (Ed.), *The Oxford Handbook of Free Will*. New york: Oxford.

Kant, I. (1887). *The Philosophy of Law: An Exposition of the Fundamental Principles of Jurisprudence as the Science of Right* (W. Hastie, Trans.). Edinburgh: T&T Clark.

Kim, J. (1984). Conceps of Supervenience. *Philosophy and Phenomenological Research, 45*(2), 153-176.

Kim, J. (2000). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2002). Multiple Realization and the Metaphysics of Reduction. In D. J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press

Kripke, S. (1972, 1980). *Naming and Nessecity*. Cambridge, MA: Harvard University Press.

Lepore, E., & Loewer, B. (2011). More on Making Mind Matter. In E. Lepore & B. Loewer (Eds.), Meaning, Mind, and Matter: Philosophical Essays. Oxford Scholarship Online: Oxford University Press. doi:10.1093/acprof:oso/9780199580781.001.0001

Levy, N. (2008). Counterfactual intervention and agents' capacities. *The Journal of Philosophy, 105*(5), 223-239.

Lewis, D. (1966). An Argument for the Identity Theory. *The Journal of Philosophy, 63*(1), 17-25.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. In D. J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings* Oxford: Oxford University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

List, C., & Menzies, P. (2009). Non-Reductive Physicalism and the Limits of the Exclusion Principle. *The Journal of Philosophy, 105*(9), 475-502.

List, C., & Menzies, P. (2017). My Brain Made Me Do It. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford Scholarship Online: Oxford University Press.

Lowe, E. J. (2000). Causal closure principles and emergentism. *Philosophy, 75*(4), 571-585.

Mackie, J. L. (1979). Mind, Brain and Causation. *Midwest Studies in Philosophy, 4*(1), 19-29.

Mackor, A. R. (2013). What Can Neuroscience Say About Responsibility? In N. A. Vincent (Ed.), *Neuroscience and Legal Responsibility*. New York: Oxford Univerisy.

McKenna, M., & Coates, J. D. (2015). Compatibilism. In E. N. Zalta (Ed.), *Stanford Encyklopedia of Philosophy* (Summer 2015 ed.).

McLaughlin, B. (2006). Is Role-Functionalism Committed to Epiphenomenalism? *Consciousness Studies, 13*(1-2), 39-66.

McLeod, O. (2013). Desert. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013 edition ed.).

McPherson, T. (2015). Supervenience in Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2015 ed.): Metaphysics Research Lab, Stanford University.

Mele, A. R. (1992). *Springs of Action*. New York: Oxford University Press.

Mele, A. R., & Robb, D. (1998). Rescuing Frankfurt-style cases. *Philosophical Review, 107*(1), 97-112.

Meynen, G. (2014). Neurolaw: Neuroscience, Ethics, and Law. Review Essay. *Ethical Theory and Moral Practice*.

Meynen, G. (2016). Neurolaw: recognizing opportunities and challenges for psychiatry. *Journal of Psychiatry and Neuroscience, 41*(3-5).

Moore, M. S. (1993). Justifying Retributivism. *Israel Law Review, 27*(1-2), 15-49.

Moore, M. S. (2016). Stephen Morse on the Fundamental Psycho-Legal Error. *Criminal Law and Philosophy, 10*(1), 45-89.

Morse, S. J. (2004). New Neuroscience, Old problems. In B. Garland (Ed.), *Neuroscience and the Law; Brain, Mind and the Scales of Justice*. New York: Dana Press.

Morse, S. J. (2006). Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio State Journal of Criminal Law*(3), 397 - 412.

Morse, S. J. (2007). The non-problem of free will in forensic psychiatry and psychology. *Behavioral Sciences & the Law*. doi:10.1002/bsl.744

Morse, S. J. (2013a). Common Criminal Law Compatibilism. In N. A. Vincent (Ed.), *Neuroscience and Criminal Responsibility*. New York: Oxford University Press.

Morse, S. J. (2013b). Responsibility and Mental Capacity. In N. A. Vincent (Ed.), *Neuroscience and Legal Responsibility* (pp. 27-52). New York: Oxford University Press.

Murray, D., & Nahmias, E. (2012). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*.

Nahmias, E., Coates, J. D., & Kvaran, t. (2007). Free Will, Moral Responsibility, and Mechanisms: Experiments on Folk Intuitions. *Midwest Studies in Philosophy, 31*(1), 214-242.

Nelkin, D. K. (2011). *Making Sense of Freedom and Moral Responsibility*. Oxford: Oxford University Press.

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs, 41*(4), 663-685.

Noë, A. (2010). *Why you are not your brain, and other lessons from the biology of consciousness*: Hill and Wang.

O'Connor, T., & Franklin, C. (2019). Free Will. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 ed.): Metaphysics Research Lab, Stanford University.

O'Connor, T. (2000). Causality, Mind, and Free Will. *Philosophical Perspectives: Action and Freedom, 14*, 105-117.

Palmer, D. (2104). Free Will, Libertarianism, and Kane. In D. Palmer (Ed.), *Libertarian Free WIll: Contemporary Debates*.

Papineau, D. (1998). Mind the Gap. In J. E. Tomberlin (Ed.), *Language,Mind and Ontology* (pp. 373-388). Blackwell: Blackwell.

Pardo, M. S., & Patterson, D. (2010). Philosophical Foundations of Law and Neuroscience. *University of Illinois Laws Review*(4).

Pardo, M. S., & Patterson, D. (2013). *Minds, Brains and Law: The Conceptual Foundations of Law and Neuroscience*. New York: Oxford University Press.

Pereboom, D. (2001). *Living without Free Will* New York: Cambridge University Press.

Pereboom, D. (2002). Living without free will: the case for hard incompatibilism. In R. Kane (Ed.), *The Oxford Handbook of free will*. New York: Oxford University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Petoft, A. (2015). Neurolaw: A brief introduction. *Iranian Journal of Neurology, 14*(5), 53-58.

Place, U. T. (1956). Is Consciousness a Brain Process? . *British Journal of Psychology, 47*, 44-50.

Plunkett, D., & Sundell, T. (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers Imprint, 13*(23).

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D. D. Merril (Eds.), *Art, Mind, and Religion* (pp. 37-48). Pittsburgh: University of Pittsburgh Press.

Putnam, H. (1975a). Brains and Behavior. In H. Putnam (Ed.), *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, H. (1975b). Minds and Machines. In H. Putnam (Ed.), *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Ramsey, W. (2019). Eliminative Materialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.): Metaphysics Research Lab, Stanford University.

Rawls, J. (1955). Two Concepts of Rules. *Philosophical Review, 64*(1), 3-32.

Robinson, H. (2017). Dualism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.): Metaphysics Research Lab, Stanford University.

Roskies, A. L., & Nichols, S. (2008). Bringing moral responsiiblity down to earth. *Journal of philosophy, 105*, 371-388.

Russell, P. (1990). Hume on Responsiility and Punishment. *Canadian Journal of Philosophy, 20*(4), 539-563.

Russell, P. (2008). Strawson's Way of Naturalizing Responsibility. In M. McKenna & P. Russell (Eds.), *Free Will and Reactive Attitudes (2008)* Ashgate Publishing Limited.

Sapolsky, R. (2006). The frontal cortex and the criminal justice system. In S. Zeki & O. Goodenough (Eds.), *Law and the Brain*. Oxford: Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In H. Feigl & M. Scriven (Eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis* (Vol. Minnesota Studies

in the Philosophy of Science (Volume 1, pp. 253–329). Minneapolis: University of Minnesota Press.
Sidgwick, H. (1907). *The Methods of Ethics* (7th ed.). Indianapolis: Hacket.
Skinner, B. F. (1948). (1962 ed.). New York: MacMillan.
Smart, J. J. C. (1959). Sensations and Brain Processes. *Philosophical Review, 68*, 141-156.
Smart, J. J. C. (1961). Free Will, Praise, and Blame. *Mind, 70*(279), 291-306.
Smart, J. J. C. (2017). The Mind/Brain Identity Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.): Metaphysics Research Lab, Stanford University.
Smilansky, S. (2000). *Free Will and Illusion*. Oxford: Oxford University Press.
Smith, M. (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In S. Stroud & C. Tappolet (Eds.), *Weakness of Will and Practical Irrationality*. Oxford Scholarship Online: Oxford University Press.
Smith, M. (2004). *Ethics and the A Priori. Selected Essays on Moral Psychology and Meta-Ethics*. . Cambridge: Cambridge University Press.
Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience, 14*(7), 513-518.
Strawson, P. F. (1962a). Freedom and Resentment. *Proceedings of the British Academy, 48*, 1-25.
Strawson, P. F. (1962b). Freedom and Resentment. In G. Watson (Ed.), *Free Will*. New York: Oxford University Press.
Taylor, R. (1960). *Action an Purpose*: Prentice Hall.
van de Poel, I. (2011). The Relation Between Forward-Looking and Backward-Looking Responsibility. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral Responsibility: Beyond Free Will and Determinism*. Springer.
van Inwagen, P. (1975). The Incompatibility of Free Will and Determinism. *Philosophical studies, 27*, 185-199.
van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

van Inwagen, P. (2015). Some Thought on An Essay on Free Will. *The Harvard Review of Philosophy, XXII*.

Vincent, N. A. (2009a). On the Relevance of Neuroscience to Criminal responsibility. *Crim Law and Philos*. doi:10.1007/s11572-009-9087-4

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In *Moral Responsibility* (pp. 15-35): Springer.

Vincent, N. A. (2013). Blame, desert and compatibilist capacity: a diachronic account of moderateness in regards to reasons-responsiveness. *Philosophical Explorations, 16*(2), 178-194.

Walen, A. (2016). Retributive Justice. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2016 ed.): Metaphysics Research Lab, Stanford University.

Walter, H. (2001). *Neurophilosophy of free will: From libertanian illusions to a concept of natural autonomy*.

Watson, G. (1987). Free Action and Free Will. *Mind, 96*(382), 145-172.

White, S. (2007). Property Dualims, Phenomenal Concepts, and the Semantic Premise. In T. Alter & S. Walter (Eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Widerker, D. (1995a). Frankfurt's Attack on the Principle of Alternate Possibilities. *Philosophical Review, 104*, 247-261.

Widerker, D. (1995b). Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities. *Philosophical Review, 104*(2), 247-261.

Widerker, D. (2000). Frankfurt's Attack on the Principle of Alternative Possibilities: a Further Look. *Philosophical Perspectives, 34*(14), 181-201.

Widerker, D. (2006). Libertarianism and the Philosophical Significance of Frankfurt Scenarios. *Journal of philosophy, 103*(4), 163-187.

Wilson, J. (1999). How Superduper Does a Physicalist Supervenience Need to Be? *Philosophical Quarterly, 49*, 33-52.

Yablo, S. (1992). Mental Causation. *The Philosophical Review, 101*(2), 245-280.

Yablo, S. (2008). *Thoughts: Papers on Mind, Meaning, Modality*. Oxford: Oxford University Press.

Yalowitz, S. (2014). In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014 ed.): Metaphysics Research Lab, Stanford University.

Young, J. Z. (1978). *Programs of the Brain*. Oxford: Oxford University Press.

ACTA PHILOSOPHICA GOTHOBURGENSIA
ISSN 0283-2380

Editors: Bengt Brülde, Ali Enayat, Anna-Sofia Maurin, and Christian Munthe

Subscriptions to the series and orders for individual copies are sent to:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, 405 30 Göteborg, Sweden
acta@ub.gu.se

Volumes published:

1. MATS FURBERG, THOMAS WETTERSTRÖM & CLAES ÅBERG (eds.). *Logic and abstraction: Essays dedicated to Per Lindström on his fiftieth birthday*. Göteborg 1986

2. STAFFAN CARLSHAMRE. *Language and time: An attempt to arrest the thought of Jacques Derrida.* Göteborg 1986

3. CLAES ÅBERG (ed.). *Cum grano salis: Essays dedicated to Dick A. R. Haglund.* Göteborg 1989

4. ANDERS TOLLAND. *Epistemological relativism and relativistic epistemology*. Göteborg 1991

5. CLAES STRANNEGÅRD. *Arithmetical realizations of modal formulas*. Göteborg 1997

6. BENGT BRULDE. *The Human Good.* Göteborg 1998

7. EVA MARK. *Självbilder och jagkonstitution*. Göteborg 1998

8. MAY TORSETH. *Legitimate and Illegitimate Paternalism in Polyethnic Conflicts*. Göteborg 1999

9. CHRISTIAN MUNTHE. *Pure Selection: The Ethics of Preimplantation Genetic Diagnosis and Choosing Children without Abortion*. Göteborg 1999

10. JOHAN MÅRTENSSON. *Subjunctive Conditionals and Time: A Defense of a Weak Classical Approach*. Göteborg 1999

11. CLAUDIO M. TAMBURRINI. *The "Hand of God"? Essays in the Philosophy of Sports*. Göteborg 2000

12. LARS SANDMAN. *A Good Death: On the Value of Death and Dying*. Göteborg 2001

13. KENT GUSTAVSSON. *Emergent Consciousness: Themes in C.D. Broad's Philosophy of Mind*. Göteborg 2002

14. FRANK LORENTZON *Fri Vilja?* Göteborg 2002

15. JAN LIF. *Can a Consequentialist be a real friend? (Who cares?)* Göteborg 2003

16. FREDRIK SUNDQVIST. *Perceptual Dynamics: Theoretical foundations and philosophical implications of gestalt psychology*. Göteborg 2003

17. JONAS GREN. *Applying utilitarianism: The problem of practical action-guidance*. Göteborg 2004

18. NIKLAS JUTH. *Genetic Information Values and Rights: The Morality of Presymptomatic Genetic Testing.* Göteborg 2005

19. SUSANNA RADOVIC. *Introspecting Representations*. Göteborg 2005

20. PETRA ANDERSSON. *Humanity and nature: Towards a consistent holistic environmental ethics*. Göteborg 2007

21. JAN ALMÄNG. *Intentionality and intersubjectivitiy*. Göteborg 2007

22. ALEXANDER ALMÉR. *Naturalising intentionality: Inquiries into realism & relativism*. Göteborg 2007

23. KRISTOFFER AHLSTRÖM. *Constructive Analysis: A Study in Epistemological Methodology*. Göteborg 2007

24. RAGNAR FRANCÉN. *Metaethical Relativism: Against the Single Analysis Assumption*. Göteborg 2007

25. JOAKIM SANDBERG. *The Ethics of Investing: Making Money or Making a Difference?* Göteborg 2008

26. CHRISTER SVENNERLIND. *Moderate Nominalism and Moderate Realism*. Göteborg 2008

27. JÖRGEN SJÖGREN. *Concept Formation in Mathematics*. Göteborg 2011

28. PETER GEORGSSON. *Metaphor and Indirect Communication in Nietzsche*. Göteborg 2014

29. MARTIN KASÅ. *Truth and Proof in the Long Run: Essays on Trial-and-Error Logics*. Göteborg 2017

30. RASMUS BLANCK. *Contributions to the Meta-mathematics of Arithmetic: Fixed Points, Independence, and Flexibility*. Göteborg 2017

31. MARTIN FILIN KARLSSON. *All There Is: On the semantics of Quantification over Absolutely Everything*. Göteborg 2018

32. PAUL KINDVALL GORBOW. *Self-Similarity in the Foundations*. Göteborg 2018

33. YLWA SJÖLIN WIRLING. *Modal Empiricism Made Difficult: An Essay in the Meta-Epistemology of Modality*. Göteborg 2019

34. MARCO TIOZZO. *Moral Disagreement and the Significance of Higher-Order Evidence*. Göteborg 2019

35. ALVA STRÅGE. *Minds, Brains, and Desert: On the relevance of neuroscience for retributive punishment*. Göteborg 2019