UNIVERSITY OF COPENHAGEN

Freedom in Uncertainty

PhD Dissertation

by

Filippos Stamatiou

February 2022

Freedom in Uncertainty: Control, Luck, and Moral Responsibility

PhD Dissertation

By Filippos Stamatiou

University of Copenhagen

Department of Communication

Section of Philosophy

Principal Supervisor: Thor Grünbaum

Secondary Supervisor: Matthew Talbert

Secondary Supervisor: Paul Russell

Word Count (excl. bibliography): 42,821

Table of Contents

Acknowledgements

Σήμερον εμού, αύριον ετέρου, ουδέποτε τινός.

It is said that philosophy is a solitary endeavour, and I could not disagree more. Both this dissertation and my philosophical journey so far have been defined by people. If what follows bears any meaning at all, it is because of them.

First and foremost, I am sincerely grateful to my supervisor, Thor Grünbaum. None of this would be possible without him. His empathy, philosophical rigour, and his willingness to engage with my ideas for the past 4 years, have almost singlehandedly turned an economics graduate with philosophical leanings into -dare I say- a philosopher. Sometimes life gives you people who are catalysts of change and personal development. Thor is exactly this person for me. I have no doubt that if every PhD supervisor was more like Thor, graduate students would be happier, more fulfilled, and philosophy would be better off as a result. Thank you, Thor.

I am also grateful to my two secondary supervisors, Matthew Talbert, and Paul Russell. Their much-needed insights and support have come at several crucial points during my PhD. Thanks also goes out to the people at the Lund-Gothenburg Responsibility Project, who have on numerous occasions given me the opportunity me to present my work. I also want to thank Sebastian Watzl and the Mind Group in Oslo, who welcomed me with open arms and actively engaged with my work, at a time when the pandemic had left most of us working in isolation.

I have been lucky in other regards, too. I come from a place of immense beauty, rich culture, deep contradictions, and endless conversation. More often than not, we deliberate for the sake of deliberation. And while this may be counterproductive in our efficiency-obsessed world, it makes for long, soulful discussions about everything that is or could be. Some call this *αμπελοφιλοσοφία* (vine-philosophy), but it has seed of deep reflection about the world in it. To my Athenian friends; Elena, Dimitris, Yiannis, Petros, Manos, Afroditi, Sotiris, Nikos, Eddie, and so many more. Thank you, I love you.

My colleagues at the University of Copenhagen have been an absolute blessing, during times thick and thin. Our lunch talks can range from advice for a zombie apocalypse to the correct pronunciation of Ancient Greek. I am thankful to so many of you; Katla and Anne-Sofie for being my friends, and for always letting me ramble on with my *semi-informed strong opinions.* Juan and Felipe for crashing student parties, listening to Colombian music and debating predictive coding at 4 a.m.. Ody for always asking how I am, and for occasionally being my winter-bathing buddy. Patricia, for this bottle of snaps I have been saving for too long, and now it's probably off. Linas for always and never leaving. Our Friday gatherings are a true delight for me.

Copenhagen has been kind to me, and I do not mean the weather. For the past 6 years, I am surrounded by people who give meaning to the term *hyggeligt*. Emma, my Danish everything, who I couldn't even begin to thank enough, but who I am most comfortable laughing with (and sometimes at). Joakim, who sometimes reminds me of myself so much it scares me. Helen, for innumerable deep and personal conversations. Angelo and Seit, who left the city, and their mark. Alicja, for showing me that partnership is easy but also hard. Tasja, for her optimistic pessimism and for her negative ape-index. Mauricio, for long talks about literature and art, on the grass or on the couch. Abe for pushing me to be better. Bartek and Mikkel, who will always be the best hosts. Lioubi, Sotiris, Xenofon, Dimitris, Chris, Evi, Christina, Alex, for being way more socially skilled than necessary.

Ultimately, I am most grateful for the love I have received and shared. My mother Dominique taught me that love is the one driving force in life. My father Yannis ignited my curiosity and excitement about the world. My stepdad Dimitris taught me the value of character, persistence, and responsibility. My grandpa Spiros ignited my love for books and ideas. My grandma Doxa provided the example of devotion and commitment. My grandma Danae is the role model for a life well-lived, driven by freedom, art, and love. Finally, Anna, for being my person.

Copenhagen, February 18, 2022

Abstract

This work develops a philosophically credible and psychologically realisable account of control that is necessary for moral responsibility. We live, think, and act in an environment of subjective uncertainty and limited information. As a result, our decisions and actions are influenced by factors beyond our control. Our ability to act freely is restricted by uncertainty, ignorance, and luck. Through three articles, I develop a naturalistic theory of control for action as a process of error minimisation that extends over time. Thus conceived, control can serve to minimise the influence of luck on action and enable freedom in uncertainty.

In Article 1, Thor Grünbaum and I argue for a psychologically plausible account of the kind of control that is necessary for moral responsibility. We begin by establishing the relationship between *agentive contribution* and *responsibility-level control*. One way to determine whether one the right kind and degree of control to be morally responsible is to track one's degree of contribution to an action. However, a psychologically plausible account of control in terms of agentive contribution may seem to contradict the types of functional-mechanistic explanations used in cognitive science. In cognitive psychology and cognitive neuroscience, personal-level capacities are explained by a set of sub-personal mechanisms. Often, such explanations leave no room for a contribution by the agent. By integrating insights from theories of cognitive control and incorporating them into a philosophical account of intentions, we propose a way of thinking about the distribution of cognitive control resources as something the agent does.

Article 2 argues that a classic argument concerning luck, originally aimed at libertarianism, generalises beyond any specific theory of free will, and regardless of whether determinism or indeterminism is true. I call this *mental luck*. Because we all make decisions under conditions of relative uncertainty and limited information, it is possible for an agent to make decisions that are contrary to their own motivation. In such situations, it may be a matter of luck whether the agent makes the right decision. Mentally lucky decisions are not rationally governed by attitudes within the agent's perspective, and thus, may be indistinguishable from unlucky ones. From the perspective of the agent, such decisions may resemble the outcome of a lottery. Therefore, mental luck poses a challenge to most prominent theories of free action and moral responsibility.

Finally, article 3 engages with the issue of *resultant luck*, namely luck in how things turn out. Resultant luck raises a challenge for theories of moral responsibility because its existence suggests that one may be responsible for factors beyond one's control. Prominent responses to resultant luck led to a choice between internalism and scepticism. I argue that familiar cases of resultant luck are based on the assumption that actions are events. Instead, I propose an alternative ontology of action as an ongoing goal-directed process with a *many-shots* structure. Described this way, cases of resultant luck are not representative of ordinary action. The proposal of action as a *many-shots* process is consistent with predictive coding, a cognitive architecture which centres around error minimisation. Under this framework, cases of resultant more luck are no longer failures of action, but rather anticipated errors to be settled within the ordinary process of action.

Resumé

Dette projekt udvikler filosofisk begrundet og psykologisk realiserbar teori om den kontrol, der er forudsætningen for moralsk ansvar. Vi lever med subjektiv usikkerhed og begrænset information. Som følge heraf er vores beslutninger og handlinger påvirket af faktorer udenfor vores kontrol. Vores evne til fri handling er begrænset af usikkerhed, uvidenhed og held. Gennem tre artikler udvikler jeg en naturalistisk teori om kontrol af handling som en fejlminimeringsprocess, der strækker sig over tid. Opfattet på denne måde kan kontrol tjene til at minimere helds indflydelse på handling og muliggøre frihed i usikker verden.

I artikel 1 argumenterer Thor Grünbaum og jeg for en psykologisk plausibel teori om den form for kontrol, der er nødvendig for moralsk ansvar. Vi starter med at etablere forholdet mellem agentens bidrag til handlingsprocessen og den form for kontrol, der er nødvendig for moralsk ansvar. En måde at afgøre hvorvidt en agent har den påkrævede form for kontrol er at spørge til graden af agentens bidrag til handlingsprocessen. En psykologisk plausibel teori af kontrol forstået som agentens bidrag kan synes at stride imod de typer af funktionelle-mekanistiske forklaringer, man bruger i kognitionsvidenskab. I kognitionspsykologi og kognitiv neurovidenskab forklares personens evner og kapaciteter ved et sæt af sub-personelle mekanismer. Ofte vil en sådan forklaringsform ikke efterlader nogen plads til agentens bidrag. Ved at integrere indsigter fra teorier om kognitiv kontrol og inkorporere dem i en filosofisk redegørelse for intentioner, foreslår vi en måde at tænke på fordelingen af kognitive kontrolressourcer som noget, agenten gør.

Artikel 2 argumenterer for, at en et klassisk argument angående held oprindeligt rettet mod libertarianisme, generaliserer udover enhver specifik teori om fri vilje, og uafhængigt af om determinisme eller indeterminisme er sandt. Jeg kalder dette for mentalt held. Da vi alle træffer beslutninger under forhold med relativ usikkerhed og begrænset information, er det muligt for agenten at komme til at træffe beslutninger, der strider imod deres egen motivation. I en sådan situation kan det være et spørgsmål om held, om agenten træffer den rette beslutning. Mentalt heldige beslutninger er ikke rationelt styrede af holdninger indenfor agentens perspektiv. Mentalt heldige beslutninger kan muligvis ikke skelnes fra ikke-heldige beslutninger. Set fra agentens perspektiv kan sådanne beslutninger ligne med resultatet af et lotteri. Derfor udgør mentalt held en udfordring for de fleste fremtrædende teorier om fri handling og om moralsk ansvar.

Endelig beskæftiger artikel 3 sig med spørgsmålet om resulterende held, nemlig held i måden hvorpå konsekvenser af ens handling falder ud. Det resulterende held rejser en udfordring for teorier om moralsk ansvar, fordi dets eksistens antyder, at man kan blive ansvarlig for faktorer uden for ens kontrol. Fremtrædende reaktioner på resulterende held fører til et valg mellem internalisme og skepticisme. Jeg argumenterer for, at velkendte tilfælde af resulterende held er baseret på den antagelse, at handlinger er begivenheder. Jeg foreslår i stedet en alternativ handlingsontologi, hvor handling er en løbende målrettet proces med en fejlminimeringsstruktur. Beskrevet på denne måde er tilfælde af resulterende held ikke repræsentative for almindelig handling. Handling forstået som en proces, hvor fejl minimeres over tid, er i overensstemmelse med prædiktiv kodning forstået som en generel kognitiv arkitektur.

Introduction

Preface

The work presented in this dissertation originates in keen curiosity about the mind, thought, and action. The dissertation comprises of 3 articles, each raising issues about highly specialised debates at the intersection of philosophy of mind, philosophy of action, cognitive and moral psychology. Together, the articles provide support for the claim that humans have the sort of free will necessary to be morally responsible. My proposal suggests that we can psychologically realise the kind of control that is necessary for moral responsibility, even though we act in an environment of subjective uncertainty and limited information.

As the title suggests, I defend an account of freedom in uncertainty. I take this as the sort of freedom we can realistically expect to have, given all our constraints. Because freedom is typically associated with control, I seek to justify an account of control which can at the same time be empirically credible, corroborate common intuitions about moral responsibility, and satisfy the conditions set by prominent theories of action. In the process, I discuss several challenges and objections to such an account of control and propose ways to resist each of them. Every challenge contributes to a more developed account of the sort of control that can enable freedom in uncertainty.

This introductory chapter offers an overview of the free will debate, focusing on the interplay between metaphysical discussions on the nature of free will, the importance of the notion of control, and the perniciousness of the problem of luck for theories of moral responsibility. I begin with a statement of the central problem the dissertation is addressing, and a short description of the contribution it offers to the literature. I move to introduce a few fundamental principles that mark my perspective throughout this work. Then, I provide an overview the relevant philosophical literature on free will and moral responsibility. This overview is organised in three parts. First, I review the problem of free will, sketching out my own compatibilist position which remains agnostic on the true metaphysical nature of the universe. Next, I discuss control, a key notion in the moral responsibility debate that spans philosophy and psychology. Here, I maintain that independently of which position one assumes in the debate, the relevant notion of control should be realisable by human psychology. Third, I turn to the problem and review two ways that factors outside an agent's control appear to make a moral difference. I briefly introduce the relevant

literature before I outline my own position in the luck debate, as presented in articles 2 and 3. The purpose of this theoretical overview is twofold: first, to introduce important concepts and debates from the literature and second, to motivate my position and sketch out how I argue for it in the three articles that follow.

I then turn to a number of methodological considerations that have guided my work. With both the theoretical background and the methodological discussion complete, I then outline each of three articles that make up my dissertation. My aim here is not only to present each article in isolation, but to draw connections between them and substantiate how they jointly constitute a coherent proposal. At the end of this introductory chapter, I discuss a common narrative which emerges from the three articles that follow.

Problem Statement

Is there a philosophically plausible and psychologically realisable account of the sort of control necessary for free will and moral responsibility, given the fact that humans think and act within an environment of subjective uncertainty and limited information?

Guiding Principles

At this point, I want to discuss three fundamental premises that underpin the work presented in this dissertation. They take the form of guiding principles for the sort of account I want to develop. Together, they make up the foundation of my proposal, the foundation of the sort of free will one can have within an environment of uncertainty.

First, throughout the work presented here, I propose to turn the spotlight on the agent. I take this as the first step towards a plausible account of morally responsible action. Once we focus on how things look from the perspective of an agent, it becomes clear that whatever the conditions form moral responsibility end up being, they should be realisable by agents like us. Furthermore, focusing on the subjective outlook of the agent underscores the importance of taking both the psychological and the epistemic capacities and limitations of ordinary agents into account. For instance, any good account of moral responsibility should be consistent with what we know about how humans exercise control over their environment. This first assumption comes into play in article 1, where Thor Grünbaum and I consider whether the sort of control that is commonly

regarded as necessary for moral responsibility is in fact realisable by human psychology. In our attempt to provide a mechanistic explanation of our capacity for responsibility-level control, we argue for the importance of agentive contribution in determining that a happening is in fact an action. The same focus on the agent underlies the arguments made in the remaining two articles.

Second, irrespective of the objective nature of the universe, we are all faced with uncertainty and incomplete information. This second premise suggests that from the perspective of the agent, whether the universe is deterministic or indeterministic has little bearing on how she acts and makes decisions. Ordinary agents are epistemically limited, in a way that precludes them from having access to complete information about the world. Within this environment of uncertainty, explaining the cognitive process of exercising control using limited resources becomes the focal point of my investigation. With the premise of perspectival uncertainty and limited information as a fundamental starting point, some familiar issues in the free will debate appear differently than they are typically portrayed in the literature. This forms the backbone of the arguments presented in articles 2 and 3. It means that the metaphysical debate about determinism and indeterminism is much less pertinent for some questions about moral responsibility. As long as the agent acts within an environment of uncertainty, it makes no difference to her whether the universe is deterministic or not. Note that this does not mean the metaphysical debate is no longer interesting in general. Rather, it means that my arguments are independent to the truth of determinism.

Finally, I come to the third guiding principle. This is the idea that action is a process of many shots. It runs contrary to the traditional event ontology of action which has dominated the field. According to the *many-shots* structure, ordinary action is best understood as an ongoing, goal-directed process that extends over time. Furthermore, this principle allows that the agent may fail and try again, while still be performing the same action. In fact, errors and misses appear less as conclusive failures of action, and more as constitutive parts of a process of constant refinement and error minimisation. Coupled with a cognitive architecture such as predictive coding, the view of action as a many-shots process has important implications for various debates within the free will literature, such as resultant moral luck. While the many-shots process ontology of action is developed in the third article, it underlies the entirety of this work, as a credible alternative to the traditional understanding of action.

Theoretical Overview

In what follows, I provide an introduction to the free will and moral responsibility literature, as well as key concepts from philosophy of action. This overview of relevant theoretical background is in no way complete. The fields I engage with are both several and individually broad. To exhaust these topics would require one or several handbooks, of which many exist. However, I provide a brief overview of the concepts and debates that are most important for the work presented in this dissertation. The aim of this presentation is to throw some light on how these concepts and debates could be informed by adopting my three basic premises. So, throughout this synopsis of the state of the art I attempt to convey how each concept, theory, or debate, is relevant for my purposes.

The Free Will Debate

Free Will

In this first part of the theoretical overview, I introduce the concepts of free will and moral responsibility before I outline the problem of free will. After reviewing the two leading camps in the debate, compatibilism and incompatibilism, I discuss their respective commitments regarding the nature of the universe. Finally, I sketch out my own position as a version of compatibilism and consider its place within the free will debate.

One could be excused for thinking that the definition free will invites little debate. After all, the idea of a free will seems both natural and quite straightforward. However, there are several different ways to define free will. How one defines free will matters, because it partly determines the sorts of answers one can give to important questions about free will. Consider three distinct ways to define free will:

1. Free will as access to alternative possibilities (van Inwagen, 1983: 8; Ginet, 1990: 90; Clarke, 2003:3)
2. Free will as the power to be the ultimate creator and sustainer of one's ends and purposes (Kane, 1996: 4)
3. Free will as the ability to exercise the kind of control necessary for moral responsibility (Mele, 2006: 17; Haji, 2009: 18)

For the work presented in this dissertation, it is the third way of defining free will that is of interest to me. The connection between free will and moral responsibility is central to the arguments I develop in the three articles that follow, and to the common narrative that emerges from them. Broadly understood, my aim is to provide credibility for the view that the ability to exercise the kind of control necessary for moral responsibility is within the range of ordinary human agents and can in principle be realised by human psychology. In turn, a credible account of this kind of control provides support for the view that some humans sometimes have free will.

Moral Responsibility

Settled on a definition of free will that centres around an ability necessary for moral responsibility, I should spell out moral responsibility in more detail. As it is used throughout this work, moral responsibility refers to the degree to which a person is worthy of blame or praise. In other words, to say that one is morally responsible is to say that they are blameworthy or praiseworthy. This definition of moral responsibility distinguishes it from other related concepts. For instance, moral responsibility is distinct from morality in general. The latter may refer to moral obligations, prohibitions, considerations of right and wrong, virtue and vice, good and bad. Moral responsibility is only concerned with the degree to which a person is worthy of blame or praise. Moreover, there is a marked contrast between moral responsibility and other types of responsibility, such as causal responsibility, legal responsibility, or aesthetic responsibility. These and other types of responsibility are not exclusively concerned solely with blameworthiness and praiseworthiness.

Before moving on, consider the concepts of blameworthiness and praiseworthiness. They are perhaps the central defining concepts of moral responsibility, which is why they feature in its definition. In essence, blameworthiness and praiseworthiness are two poles on the moral responsibility continuum (Khoury, 2018). To be either praiseworthy or blameworthy entails being morally responsible. Or one may conceive them as the two only types of responsibility. Then, being morally responsible means either being blameworthy or being praiseworthy. Regardless of which understanding of blameworthiness and praiseworthiness one chooses, the important point is this: moral responsibility refers to the degree to which someone is worthy of blame or praise.

Reactive Attitudes

But what does it mean exactly to be worthy of blame and praise? The answer to this question will significantly illuminate our understanding of the kind of moral responsibility we are interested

in. According to P.F. Strawson (1962) and arguably most philosophers since, blaming and praising people for their actions involves certain affective responses, namely *reactive attitudes*. Strawson's focus on reactive attitudes reflects the social significance of the practice of holding each other responsible. Reactive attitudes typically include resentment, gratitude, forgiveness, anger, love, and indignation, among others.

Strawson made an important distinction between *detached* or *objective* attitudes on the one hand, and *non-detached* or *reactive* attitudes on the other. Assuming the objective attitude towards another person involves being detached from them in some way. Certain responses, such as resentment, indignation or forgiveness are less appropriate, because the person is treated partly as an object, part of the causal chain but not in the right way to be truly worthy of blame and praise. Everyday examples of assuming the objective attitude with someone typically affect the degree to which that person is blameworthy or praiseworthy. Conversely, assuming the reactive attitude towards someone is often taken as the prototypical example of engaged human interaction. For many philosophers of moral responsibility, being a morally responsible agent is the same as being an appropriate target of the reactive attitudes, and the very concept of moral responsibility consists in extending reactive attitudes towards each other (Haji, 2002: 204).

Two considerations are relevant when it comes to Strawson's account of reactive attitudes. First, the view has both descriptive and normative aspects. On the one hand, the idea of reactive attitudes is meant to capture how we actually feel towards others, and the way we actually express moral judgement. On the other hand, their existence involves a normative demand in interpersonal relationships, specifically a moral obligation to act with good will. As Strawson notes, reactive attitudes are "natural human reactions to the good or ill will or indifference of others towards us as displayed in their attitudes and actions" (Strawson, 1962: 67). For instance, being at the receiving end of an attitude of indignation may reasonably motivate one to abstain from behaviour that could incite similar responses in the future. Second, reactive attitudes can function as both backward and forward-looking responses. For instance, when one expresses gratitude for a kind act, one is engaging a backward-looking reactive attitude of praise towards someone because of her actions in the past. At the same time, an expression of gratitude indicates how one wishes or expects to be treated in the future. Correspondingly for blame-involving responses, an expression of indignation clearly indicates that one does not wish to be treated this way, thus creating a forward-looking demand for the target of such a response. Bidirectional reactive attitudes create trajectories of reactive exchange (McGeer 2012; 10), directed both towards the past and towards the future.

Senses of Responsibility

Finally, to round up the discussion of the relevant sense of moral responsibility, there are various refinements to the concept which arguably help to capture people's intuitions and everyday moral practice better. For instance, some distinguish between responsibility as *attributability*, where something can be correctly attributed to an agent, and responsibility as *accountability*, where blameworthiness and praiseworthiness are fully operative (Watson, 1996). Others make the further conceptual distinction of responsibility as *answerability*, where it is justified to ask the responsible agent to cite reasons for something being the case (Shoemaker, 2011). Yet others insist that the relevant sense of responsibility is captured by this notion of answerability alone (Smith, 2015). For my purposes, the outcome of this debate is not important. That said, I do subscribe to a view that accepts some kind of control as a necessary condition for moral responsibility and insists that the right sort of control involves a contribution by the agent. Furthermore, my definition of moral responsibility focuses on blameworthiness and praiseworthiness. Therefore, responsibility as accountability is the most proximate sense for the work presented in this dissertation.

Moral responsibility is the central concept here because it motivates the significance of the problem of free will. We care whether we have free will, because free will justifies our holding people responsible. At the same time, to justify holding people responsible, it should be possible for them to fulfil the relevant conditions. So, whatever the conditions for someone to be morally responsible end up being, those conditions should be within the range of human psychology. That is, it should be possible for ordinary humans to fulfil whatever is required for being worthy of blame and praise. My work is an attempt to investigate the relevant limitations we face, and consider whether given those limitations, there is hope for a psychologically plausible and philosophically credible account of moral responsibility.

The Problem of Free Will

Now we can venture an articulation of the problem of free will. The problem is typically motivated by considering the truth of determinism. Intuitively, if the universe is deterministic, our capacity for control seems to be under serious threat. Before we see why, we need to define determinism. In a broad sense, determinism is the thesis that at any time only one future is physically possible (van Inwagen, 1983: 3; Kane, 2002: 27-28; Mele, 2006: 3). While there are many possible elaborations of this underspecified definition of determinism, it is adequate for my

purpose here. Determinism is a general metaphysical thesis about the nature of the universe. The thesis of determinism provides the most common introduction into the problem of free will. Consider the following:

1. Determinism is true.
2. At least some persons have free will.
3. Free will is incompatible with determinism

Compatibilism and Incompatibilism

This first formulation of the problem of free will consists of a set of mutually inconsistent propositions. Typically, framing the problem begins with accepting or rejecting proposition (3) from above, by means of posing the *compatibility question* (Kane, 1996: 13), namely whether free will is compatible with determinism. Two large camps make up the debate, compatibilists and incompatibilists. Broadly, compatibilists believe that free will and moral responsibility are compatible with determinism, while incompatibilists believe that free will and moral responsibility are incompatible with determinism (Mele, 2006: 3). These two principal positions aside, there are also several variations on them, some of which are relevant for my work. I will discuss these variations after introducing the compatibilism/incompatibilism debate in earnest.

Compatibilists and incompatibilists represent opposing views on the relationship between free will and determinism, and subsequently, different accounts of what it means to be morally responsible. In essence, the two positions emerge from a metaphysical commitment. *Compatibilism* is the view that it is metaphysically possible that determinism is true and at least some people have free will. *Incompatibilism* is the view that it is not metaphysically possible that determinism is true and at least some people have free will. Both statements are conditional. Neither compatibilists nor incompatibilists need to commit themselves to the truth of determinism. For instance, a compatibilist and an incompatibilist alike will gladly accept evidence that determinism is not true. They are only committed to the conditional relationship between the truth of determinism, on the one hand, and the possibility of some people having free will, on the other.

Furthermore, neither compatibilists nor incompatibilists commit themselves to humans actually having free will. Rather, they commit to a metaphysical relation between the truth of determinism on the one hand, and the possibility of free will on the other. For instance, as an incompatibilist one

is committed to the view that it is not possible for people to have free will if determinism is true. But it may still be true that determinism is not true, and yet people do not have free will. Correspondingly for compatibilists, the view is committed to the compatibility between determinism and the possibility of free will, not to the truth of free will in general.

That said, being a compatibilist who does not believe in free will is indeed a bizarre position to hold. If people acting freely is compatible with both determinism and indeterminism, then we should expect some of people to be able to act freely at some point. So, the compatibilist position typically maintains that we have free will, but often leaves the nature of the universe open. The position I am sketching in the three articles that constitute this dissertation is compatibilist in that sense. I insist that at least some people sometimes have free will and argue for specifying the structures that enable this sort of free will, without specifying the nature of the universe. My view is that from the perspective of an ordinary agent acting in the world, the truth or not of determinism makes no relevant difference.

Such an agnostic view is by no means canonical. Many variations of both compatibilism and incompatibilism take a stance on the conditional about free will and the truth of determinism. More specifically, some variations of compatibilism and most versions of incompatibilism take a stance on the basic nature of the universe. For instance, a version of incompatibilism called *libertarianism* claims that determinism is false and at least some people have free will (Kane, 1996; Ekstrom, 2000; O'Connor, 2000; Clarke, 2003; Ginet, 2003; Balaguer, 2010, Steward, 2012). Free will libertarians, who form arguably the majority of incompatibilists, believe that we have free will, and that free will is incompatible with determinism. Therefore, according to libertarianism, the world must be indeterministic. Thus, providing a credible account of free action within an indeterministic universe is the principal task of libertarianism. Another incompatibilist view that takes a position regarding the nature of the universe is *hard determinism*. Hard determinism is the rarely explicitly adopted view that determinism is true, and no one has free will (for a discussion see Kane, 2002: 27-32). But not all such positions are incompatibilist. *Soft determinism,* the view that determinism is true, and some people have free will (Hobart, 1934; Edwards, 1958). Because it maintains that free will can exist in a deterministic world, soft determinism is a version of compatibilism.

Free Will Scepticism

A final family of views on free will that is relevant to discuss here is *free will scepticism.* While that involves some simplification, I label as free will scepticism all views that claim no one ever has free will, as well as views which claim that we have insufficient reason to believe that anyone ever has any free will. For instance, hard determinism is such a sceptical view, because it maintains that no one has free will. So is *hard incompatibilism*, namely the view that no one has free will, either because determinism is true, or because a kind of indeterminism is true, and that kind of indeterminism is incompatible with free will (Pereboom, 1995; 2001).

For my purposes, the important point is that free will scepticism is closely associated to moral responsibility scepticism. Moral responsibility scepticism is the view that we have insufficient reason to believe that anyone is ever morally responsible, or more simply the view that no one is ever morally responsible (Strawson, 1994; Levy, 2011: 103-106, Levy, 2014: 109-126, Pereboom, 2001: 120–22). Within the definition of free will that associates it with the kind of control necessary for moral responsibility, the link between free will and moral responsibility scepticism is clear. Throughout the work presented in this dissertation, I assume that free will scepticism entails moral responsibility scepticism.

Sceptical worries feature prominently in articles 2 and 3. In article 2, I argue that because of epistemic limitations inherent to all of us, we function in an environment of subjective uncertainty and limited access to information, even concerning one's own mental states and decision-making. Within this environment and regardless of the nature of the universe, our decisions are subject to a kind of luck. The existence of this kind of luck raises sceptical worries about whether we have free will in the first place. In article 3, I focus on a different sceptical challenge, which calls into question whether people are morally responsible for their actions and their consequences. I argue against this sceptical worry by arguing that our actions are best described as on-going processes of many shots whose task is to bring the uncertainty of the world under some degree of control.

Control

Control is an important concept in the free will and moral responsibility literature, and so it is for my own work. We already have a definition free will as whatever kind of control is necessary for moral responsibility. That definition points to the centrality of the notion of control. By and large, the reason we care about free will in the first place is because it is required for moral responsibility. In free will, we look for a way to justify the practice of holding each other responsible. In turn, moral responsibility is more often than not spelled out in terms of control. Therefore, deciding exactly what kind of control is relevant for one to be blameworthy or praiseworthy, and explaining why and how this sort of control is within the range of human capacities, is an important task.

It is not the purpose of this theoretical overview to discuss the conditions of moral responsibility in detail. In short, however, there have been historically two such conditions. Aristotle expresses them in negative terms, speaking of *force* and *ignorance* as the two fundamental excusing conditions (Aristotle, 1985: 1109b30-1111b5). In positive terms, force is often expressed in terms of *freedom*, and ignorance in terms of fulfilling some *epistemic requirement* (Fischer & Ravizza, 1998; Ginet, 2000; Mele, 2010). While not universally accepted (see Björnsson, 2017), the *freedom* condition on the one hand, and the *epistemic* condition on the other, are commonly accepted as the two necessary conditions for moral responsibility. Control is typically identified with the freedom condition. While interesting, I do not discuss the epistemic condition in depth in the work presented in this dissertation (for an in-depth discussion see Wieland & Robichaud, 2017). For now, it suffices to say that control is a necessary but not sufficient condition for moral responsibility.

However, both the control condition and the concept of control in general are highly contested. Some philosophers have argued that if determinism is true, the control condition is in fact impossible to satisfy (Strawson, 1986; Kane, 1995; Pereboom, 1995). Others have recently argued that control is in fact not a necessary condition for moral responsibility (Smith, 2008, 2015; Talbert, 2019). More broadly, even if one accepts control as necessary for being blameworthy and praiseworthy, there is little agreement as to what exactly control means. Within philosophy, there are different ways to decide what kind of control is relevant, and I discuss these ways in the remainder of this section. On a different note, in cognitive psychology, control is traditionally understood as the ability to pursue our goals when confronted with habitual or otherwise compelling behaviours (Cohen, 2017). So, in the empirical sciences that study cognition, control is discussed in

contrast to automatic or habitual behaviours, and it is typically understood as a limited cognitive capacity that is generally relatively slow and subject to interference by automatic processes (Shiffrin & Schneider, 1977; Posner & Snyder, 2004). My works attempts to bridge this gap between the control in philosophy and control as it is often understood in cognitive psychology. Establishing commonalities between different approaches to the notion is the first step in proposing a psychologically realisable and philosophically relevant notion of control.

Control and Alternate Possibilities

Perhaps the most common way to spell out the relevant kind of control for moral responsibility is in terms of an ability to do otherwise. This strategy, while common, comes with its own set of problems. More specifically, it brings into view the differences between compatibilism and incompatibilism. Traditionally, the debate on the ability to do otherwise is used to force a choice between the two positions. My strategy, both in this theoretical overview and in the articles that follow, is quite different. Instead of taking a stance on the exact structure of the right kind of control, I carve out the few similarities that a notion of control has for both compatibilists and incompatibilists. Once salient, these similarities indicate that the relevant notion of control is inextricably agent-related, and that it should be realisable by human psychology.

To find support for this set of minimal conditions, one first needs to examine the conceptual origins of the disagreement on control. Compatibilists and incompatibilists largely agree that some kind of control is necessary for moral responsibility. However, they disagree on what control means exactly. Invariably, this relates to a disagreement around the ability to do otherwise. On the incompatibilist side, most accept some version of the *consequence argument* (Wiggins, 1973; Lamb 1977; Ginet, 1980; 1996; van Inwagen, 1975; 1983). Consider the following formulation, which approximates van Inwagen's (1983) argument:

1. It is not up to us to change the events of the past or the laws of nature.
2. The events of the past and the laws of nature entail every fact of the future.
   (i.e., determinism is true).
3. Therefore, the future is not up to us.

The consequence argument assumes the truth of determinism, and then goes on to argue that if determinism is true, the future is not up to us. For the incompatibilist, the argument captures

something rudimentary about any correct account of free will: the existence of alternate possibilities. A free and morally responsible agent should have options open at the moment of choosing or deciding. If every fact of the future is already decided, there are no options, and therefore, no free will to be had. This is the first and most straightforward way in which the consequence argument draws a dividing line between compatibilism and incompatibilism. Compatibilism maintains that free will is compatible with determinism. If the truth of determinism excludes the possibility of free will, compatibilism must be wrong. The requirement for alternate possibilities is captured by the following principle:

> *The Principle of Alternate Possibilities (PAP)*: A person is morally responsible for her act only if she could have done otherwise than she does (Frankfurt, 1969).

Free will requires a capacity for control, an ability to conform the world to our desires, to influence the way things turn out. Being able to do otherwise seems to be part and parcel of the relevant capacity for control. However, compatibilists and incompatibilists disagree on what the ability to do otherwise means exactly. One can trace the origins of compatibilist conditional analysis back to Hume (Hume, 1955: 104; for a discussion see Russell, 1995) A somewhat more contemporary version suggests that an agent could have done otherwise, if it is true that, had she chosen to do otherwise, then she would have done otherwise (Moore, 1912).

The incompatibilist would find that hard to accept, since the ability to do otherwise requires alternate possibilities for the agent right now (while keeping the actual past and the laws of nature fixed). At this point, the distance between the two views seems hard to bridge. The incompatibilist will insist on demanding something that the compatibilist cannot provide, namely that I, now, given all the facts about the world and my psychological structure, can do differently from what I do. Because of the consequence argument, this ability is undermined by the truth of determinism. If determinism is true, no one can ever do otherwise and therefore, no one ever has access to alternate possibilities.

Frankfurt Cases

A common way for compatibilists to argue against the conclusion of the consequence argument is to question whether the principle of alternate possibilities is in fact credible. The most familiar such account comes from Harry Frankfurt, who demonstrated that it is possible to think of cases of

secretive coercion where an agent does not have alternate possibilities, and yet, that agent is morally responsible. Consider the following case, typically referred to as a *Frankfurt case*:

> *Frankfurt Case*: "Suppose someone – Black, let us say – wants Jones to perform a certain action. Black is prepared to go to 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 to do 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's initial preferences and inclinations, then, Black will have his way". (Frankfurt, 1969: 835)

It is not important how Black ensures that Jones will do what he wants him to do. Perhaps he is a malicious neuroscientist who has implanted a device in Jones' brain, or something else to that effect. The point is that Jones has no alternate possibilities: he can only do what Black wants him to do. And yet, if the device is never triggered and Jones makes up his mind on his own and does what Black wanted him to do all along, he seems morally responsible for doing so.

For my purposes, it is also not important that Frankfurt-style cases are counterexamples to PAP. What is important is that because of such cases, one can remain open to the question of whether alternate possibilities are necessary for moral responsibility, and consequently for control. That means that we can discuss the conditions for moral responsibility independently of the consequence argument. Since my aim is to spell out minimal conditions for control that both compatibilists and incompatibilists can accept, I can remain agnostic as to whether the relevant notion of control requires alternate possibilities or not.

New Compatibilism and Control

More recently, compatibilists have argued that free will and the relevant notion of control is not captured by the incompatibilist ability to do otherwise in the first place, in what has been dubbed *new compatibilism* (Russell, 2002). For instance, Dennett claims that the kind of freedom we should want and care about does not require a general ability to do otherwise (Dennett, 1984: 2). Rather, it involves a more restricted and focused capacity for *self-control* (ibid.: 52). Arguably the most

influential and refined such account is that of *guidance control* (Fischer and Ravizza, 2000). Guidance control refers to a capacity that humans have, enabled by a reasons-responsive mechanism. That mechanism needs to fulfil two criteria:

1. Regularly receptive to reasons, some of which are moral reasons;
2. At least weakly reactive to reasons (but not necessarily moral reasons).

For Fischer and Ravizza, guidance control is all the control an agent needs to be morally responsible, and it does not require an ability to do otherwise. What it requires is that the actual causal sequence leading up to the action displays a modal property such that, under different circumstances, the agent's moderately reasons-responsive mechanism would have reacted to relevant reasons. So, while guidance control involves a certain counterfactual capacity, it does not require an ability to do otherwise, only a certain counterfactual sensitivity to reasons.

Because it is a compatibilist notion, guidance control attempts to explain how moral responsibility is compatible with the truth of determinism. It does so in the following way. Even if determinism is true and the agent couldn't have done otherwise, her moderately reasons-responsive mechanism provides a degree of counterfactual sensitivity to reasons. That sensitivity is enough for guidance control, which in turn is all the control one needs to be morally responsible. An agent with guidance control is morally responsible, even if she couldn't have done otherwise. So, guidance control does not explain how the agent could have done otherwise in the actual sequence. Rather, it is a compatibilist account of the agent's ability to act because she chose to act.

Rounding up the discussion of the different conceptualisation of control and alternate possibilities. It is not my aim to settle the debate. However, what emerges from the discussion so far is that the relevant notion of control should be recognisable in discussions of free will and moral responsibility. Apart from that, control should be referring to a set of capacities that can be realised by human psychology. To understand what sorts of capacities those may be, we need to turn to philosophy of action, and consider the importance of agenthood and agentive contribution. Some notion of agenthood is important for theories of action. But agenthood is also what binds together control in philosophical discussions about moral responsibility on the one hand, and psychological accounts of control of action on the other.

Agenthood and Action

It is often said that the task of philosophy of action is to distinguish between "what an agent does and what merely happens to him" (Frankfurt, 1978: 157). But, while making that distinction lies at the heart of any theory of action, the line between action and non-action is far from clear. Causal theories attempt to explain the difference between action and non-action by focusing on causal histories. If an event has the right causal history, it counts as an action. If not, then the event is a mere happening. However, a closer look into what are necessary or sufficient conditions for events to count as actions shows that the distinction is hard to pin down.

Consider the difference between saying that "A happened" and saying that "Smith did A". Suppose Smith decides to kill the mayor by shooting him with a gun. But having never committed homicide before, when Smith takes out the gun he starts sweating incessantly, his finger slips on the trigger and the bullet ends up killing the mayor. This is a typical case of a *deviant causal chain*, and it readily demonstrates why it may be wrong to describe something as an action, even if it aligns with the agents desires or intentions.

Some philosophers have thought that cases of deviant causal chains point to a distinction between events that are caused by other events and events that are caused by agents – that is, between event-causation and agent-causation. *Event-causal* accounts (for instance Ekstrom, 2000; also, Mele 1995; 1996) capture many people's largely physicalist intuitions about causation and agency. Trouble appears, however, when one attempts to provide an event-causal account of human action. Many philosophers assume that mental events such as beliefs, desires, or intentions are no different from other events in the physical world. But if mental events have causal powers and cause action, and they are like other events, what is the difference between the agent's actions and things that simply happen to the agent?

Agent-causal accounts (Chisholm, 1966; Clarke, 1993; O'Connor, 1996) attempt to tackle the problem by arguing that action is demarcated by a special kind of causation, namely agent-causation. An event counts as an action only if the event is attributable to the agent in a special way. Few philosophers are convinced by the metaphysics required by such accounts. For our purposes, it suffices to say that regardless of whether they succeed or not, agent-causal accounts recognise the centrality of agenthood in distinguishing actions from non-actions.

So, distinguishing between action and non-action is not always straightforward. One might be tempted to conclude that the event-causal account does not have the resources to make that distinction. So, some causal theories of action attempt to explain the special role of the agent by introducing agent-causation. There are good reasons for remaining sceptical about whether that strategy achieves its goal, but the important point is clear: Agent-causation is an attempt to fix a problem that confronts all causal accounts of action, namely distinguishing action from non-action. To make that distinction we need to posit the agent as part of the causal picture. Common intuitions point to agenthood being an integral part of any credible explanation of action.

In the context of the work presented in this dissertation two things are important: all causal theories should distinguish between action and non-action, and the line that separates them is not always clear. I have been working under the assumption that some causal account of action is the correct one. However, I do not engage in deciding which one is best. Rather, I am concerned with showing that the distinction between action and non-action is closely associated with agenthood, and that it can be spelled out in terms of the agent's contribution. In article 1, Thor Grünbaum and I spell out control in terms of a contribution by the agent, before we propose an account of the capacity for control in terms of standing intentions as structuring causes of action. Tracking the agent's contribution is a way to determine whether the agent has the right sort of control to be morally responsible.

Summing up the relevance of the debate on control. Standard forms of both compatibilist and incompatibilist accounts involve some notion of control. They may disagree on what control means exactly, but some kind of control is a requirement for moral responsibility. My contribution to the control debate can be captured in the following way: it begins with arguing that independently of one's position in the free will debate, it is still important that control is realisable by human psychology. Then, it substantiates that claim by providing a plausible account of this sort of control. Finally, it concludes by defending the plausibility of our capacity for control against sceptical challenges, most prominently luck.

Luck

The third and final concept I want to discuss in this theoretical overview is luck. In philosophy, luck is usually associated with ethics (moral luck), epistemology (epistemic luck), and political philosophy (just dessert). For the purposes of this dissertation, I am primarily interested in moral luck. Moral luck is the version of luck which affects an agent's blameworthiness and praiseworthiness. Articles 2 and 3 are dealing with different luck-related challenges to theories of free will and moral responsibility. On a secondary note, article 2 also engages with the discussion on epistemic luck. However, my focus is squarely on luck that appears to influence moral responsibility.

Luck is an elusive concept to define. Many define it in relation to control, more specifically *lack of control* (Nagel, 1979: 26; Williams, 1981: 22; for a recent review see Lang, 2021: 1-25). According to this view, luck refers to factors that are beyond one's control. Others claim that luck involves *chanciness* (see Latus, 2003), in that the occurrence of a lucky event is indeterminate, or an accident. Yet others focus on the fact that lucky events should be *significant* to the agent. For instance, the *modal account* of luck involves only two conditions, chanciness, and significance (Pritchard, 2005). Prichard's modal account of luck has been influential in epistemology. But arguably the most common way to define the general notion of luck is to associate lucky events with lack of control and significance, often dropping the chanciness condition altogether (for instance Levy, 2011: 36). While the debate on the general notion of luck is interesting, in the moral responsibility literature luck is tied to lack of control. So, throughout this dissertation, the sort of luck I am interested in is moral luck.

Moral luck

Traditionally, introductions to moral luck begin with Nagel's (1979: 24-38) taxonomy, which include four distinct kinds of moral luck. The first of these is *resultant* or *outcome luck.* Resultant luck refers to luck in the way things turn out, and it is arguably the most commonly discussed kind of moral luck in the literature. I spend the first half of article 3 discussing resultant luck and the challenge it poses to theories of moral responsibility. Another type of moral luck is *circumstantial luck,* which is luck in the morally significant circumstances an agent encounters. *Constitutive luck* is luck in the way one is constituted and may refer among other things to a person's character, natural dispositions, or traits. Finally, *causal luck* refers to "how one is determined by antecedent

circumstances" (Nagel 1979: 28). Causal luck closely resembles the implications of the truth of determinism.

Moral luck exposes an apparent conflict between our deeply held views about morality and the way things actually happen in the world. On the one hand, there seems to be a natural intuition that people are not blameworthy or praiseworthy for what is not up to them. As Nagel puts it: "Prior to reflection it is intuitively plausible that people cannot be morally assessed for what is not their fault" (Nagel. 1979: 25). This, as we have seen, is corroborated by most prominent theories of moral responsibility, in the form of the control condition. So, if the lack of control account of luck is correct, and if moral responsibility requires control, luck should not affect how blameworthy or praiseworthy an agent is. However, cases where that happens seem to be plentiful. Consider the following conspicuous case of apparent moral luck:

> *Drunk driving*: Two drivers willingly and freely decide to drink and then drive. Everything about the two drivers and the external world is identical, except one fact: one drives home safely while the other, by stroke of luck alone, runs over and kills a pedestrian. Had the pedestrian walked off the pavement a moment sooner or later, she would not have been run over, and the second driver would have driven home safely, same as the first one.

Different versions of essentially the same case have been a staple of the moral luck debate at least since Nagel's seminal essay. Now consider what *drunk driving* demonstrates. Both drivers are blameworthy for willingly and freely drinking and then driving. The second driver, though, also seems blameworthy for running over and killing the pedestrian. However, since the two drivers differ in nothing right up to the moment of the crash, the only difference between them seems to be a matter of luck. So, in a case like *drunk driving* luck seems to make a morally significant difference between the two drivers. Such cases demonstrate that moral luck exists.

As any challenges to theories of moral responsibility, the existence of moral luck invites a reaction. Broadly and simplifying somewhat, one can distinguish two camps. In the first, we find those who think that moral luck can sometimes influence the blameworthiness of agents (Hartman, 2017; Lang, 2021). Such accounts admit that factors beyond the agent's control may influence her degree of moral responsibility. In the second camp, we find people who maintain that luck should not -or rather, *does not*- influence how blameworthy or praiseworthy one is. Arguably, this is the

majority view, even though it takes comes in many shades (Zimmermann, 1987; 2002; Thomson, 1989; Latus, 2000; Khoury, 2018). I discuss the differences between the different accounts, and the potential each of them shows, in article 3. What is important to convey here is that to defend the claim that one is not blameworthy for what is beyond one's control involves explaining away the apparent influence of lucky factors on one's moral responsibility. If the influence of such lucky factors is mere appearance, luck can be explained away. If it is more than mere appearance, then those factors are morally significant, and agents are sometimes morally responsible for what is beyond their control.

So, there are fundamentally two ways to react to apparent cases of moral luck. One can either accept that moral luck makes a difference to blameworthiness or resist that conclusion and attempt to explain away the apparent influence of luck. For the work presented in this dissertation, I take the second route. Before moving on, however, I want to briefly consider what taking the first route may implicate. Suppose that one can be morally responsible for lucky events. This opens a discussion of a wealth ethical and metaethical issues. For instance, one may adopt an externalist view to moral practice, and look to the reactive attitudes as the means to determine what moral responsibility means. This sort of externalism can possibly be justified by a more consequentialist approach to blame and praise. Perhaps the second driver is more blameworthy because of forward-looking consequentialist reasons, i.e., we blame them more, so they serve a warning to future drivers. While compelling, such discussions are clearly beyond the scope of the present work.

Summing up, moral luck poses a serious challenge to the majority of moral responsibility theories. If control is a necessary condition for moral responsibility, one cannot be responsible for what is beyond one's control. Luck, in the general notion discussed above, consists of factors that are beyond one's control. Moral luck, more specifically, consists of factors beyond one's control that appear to make a morally significant difference. This is the essence of the conflict that moral luck raises. How can luck-related factors make a morally significant difference to one's moral responsibility, if they are beyond one's control? Either control is not in fact a necessary condition for moral responsibility, or luck-related factors do not make a morally significant difference, they just appear to be.

Both articles 2 and 3 are concentrated of different versions of the problem of luck. Article 2 spells out the basic structure of the challenge to my position in terms of mental luck. The article begins with a discussion of the worry about present luck (Mele, 2008: 58). While initially aimed at

libertarian theories of free will, I argue that a version of the worry about present luck, *mental luck,* may generalise beyond libertarianism. Once one adopts an agent-centred framework of action, within subjective uncertainty and limited information from the agent's perspective, certain worries one might have about luck in an indeterministic world might be inherited by all agents, independently of whether the universe is deterministic or indeterministic.

In article 3, I turn to one of the original kinds of moral luck, resultant luck. After close examination of apparent cases of resultant moral luck, I argue that, that far from being representative of ordinary action, such cases are underpinned by a problematic event ontology and thus, they are not representative of the structure of action as a whole. In response, my proposal allows agents to be responsible for actions and their consequences. Armed with an alternative ontology of action and a plausible, probabilistic cognitive architecture, I argue that the mind has evolved ways to control action in this uncertain environment. Our capacity to handle and minimise uncertainty holds some hope for a philosophically relevant and psychologically plausible account of morally responsible action. While this argument does not resist the challenge from resultant moral luck in its entirety, it does significantly diminish its force.

Methodology

Generally, my methodological approach is naturalistic. It adopts the view that philosophy should engage with empirical science. In the following section I discuss two specific methodological considerations which have guided my work. These are *reflective equilibrium* and *ought-implies-can*. The principle of reflective equilibrium has facilitated the process of justifying some of the claims presented in the three articles that follow. The principle of ought-implies-can has provided a guiding methodological principle which underscores the entirety of my proposal. Together, they have functioned as a sort of methodological compass. Before I say a bit more about how these methodological approaches have informed my work, I should introduce each of them.

A prominent method of moral justification, *reflective equilibrium* is commonly associated with moral and political philosophy and the work of John Rawls. However, though the term was introduced by Rawls (1971), the method has been around for much longer. Furthermore, reflective equilibrium is not limited to moral or political questions and can in principle be applicable to other areas of philosophy. I apply the methodology of reflective equilibrium in philosophy of mind and

action, in order to investigate whether there is reason to justify our belief that humans have the sort of free will that makes them morally responsible agents.

Reflective equilibrium suggests that it is possible to justify some of our beliefs by reaching a state of coherence between them. Incoherent beliefs are anathema in philosophy, and therefore, holding beliefs that do not contradict each other is a natural aim. That said, the best way to reach such a state of belief coherence is more debatable. According to reflective equilibrium, the required coherence involves striking a balance between three distinct elements that inform our judgements: First, common intuitions. For instance, we might have a strong natural intuition that some people act freely some of the time. Second, empirical evidence or background knowledge, such as widely accepted psychological frameworks about control, decision-making, or expected value. Third, theories about the world, including philosophical accounts of free will and moral responsibility.

As my general focus has been defending the view that people can realise the sort of freedom required for moral responsibility independently of whether the universe is deterministic or indeterministic, reflective equilibrium has featured heavily in justifying such a view. Specifically, it has been important to strive for a justificatory balance between intuition and theory. On the one hand, we do not want to rely too much on intuitions about what it means to be free and morally responsible. Such reliance might lead us to reject long-standing and cogent philosophical theories or compelling empirical data from psychology or neuroscience. On the other hand, it is important that our best theories about human thought and action are not contradicting strong natural intuitions about these matters. That would require rejecting beliefs and social practices which are part and parcel of being human. Striving to reach a state of reflective equilibrium between intuitions, empirical evidence, and theories allows us to create a set of coherent beliefs.

Now moving to the principle of *ought-implies-can*. Commonly associated with Kant (Kant, 1998; Stern, 2004), ought-implies-can is often regarded a minimal condition of justification for any moral theory. The principle expresses the idea that to be required to do something implies being able to do it. For instance, one can have a moral obligation to perform some action only if one is able to perform such action. It is a matter of some debate exactly what the *can* in ought-implies-can may mean, but in the most innocent interpretation, it refers to something being physically possible for an agent to carry out.

For purposes of the arguments presented in this work, ought-implies-can is an underlying methodological criterion of plausibility for any account of free and morally responsible action. In my view, there is an ought-implies-can relationship between philosophical theories of moral responsibility and the actual cognitive abilities of ordinary humans. For instance, if an agent is morally responsible if and only if she has a certain degree of control, then she should be able to realise such control. And if psychology can provide explanations about the cognitive capacities of humans, then psychology should be able to provide an explanation of the capacity for the sort of control that is necessary to be morally responsible. Ought-implies-can suggests that whatever the conditions of free and morally responsible action end up being, they should be realisable by human psychology.

In article 1, Thor Grünbaum and I make use of both methodological tools in order to argue for a mechanistic, psychologically realisable explanation of the capacity for cognitive control, in which the agent has an important contribution to make. In line with the principle of ought-implies-can, we argue that if control is necessary for moral responsibility and humans are in fact morally responsible, then we should be psychologically capable of realising such responsibility-level control. In our attempt to provide such a mechanistic account of responsibility-level control, we strive for a reflective equilibrium between prominent theories of action, natural intuitions about moral responsibility, and influential paradigms from cognitive psychology.

Ought-implies-can also features in article 2. Article 2 raises a worry against the common-sense view that luck should not affect the blameworthiness of agents and the evidence to the contrary. Luck may cause the outcome of a decision to be different without the agent doing anything differently, and because of epistemic limitations that we all face, all the time. This worry suggests that any decision may be prone to luck, and therefore, people do not fulfil the requirements for moral responsibility. If we are unable to ensure that a decision is immune from luck, we cannot justifiably hold a person responsible for it.

Finally, in article 3, I make use of both methodological tools to argue for an ontology of action which minimises our exposure to the negative effects of resultant moral luck. Attempting to reach a reflective equilibrium between the natural intuition that people should not be more blameworthy for things that are beyond their control and the influential view that actions are event-like completed particulars, I end up adopting an alternative process ontology. According to it, actions are ongoing processes of goal-directed behaviour that can be described as a *many-shots* process. My claim is

bolstered by a plausible cognitive framework which centres around error minimisation as the primary task of the mind. If correct, this proposal allows action to be understood as a process where luck is not a failure, but an error to be minimised. Most importantly, it suggests that people can be expected to satisfy the conditions for moral responsibility, within our limited perspective of the world.

Article Outline

This dissertation consists of three independent but closely interrelated articles. All three of them deal with current issues in the debate about free will, action, and moral responsibility. One way to approach the three articles would be to say that the first one is concerned with control, while the latter two with luck. While accurate, such a description would miss the deep connections between the issues raised in each article, and the common narrative about freedom in the face of uncertainty which emerges once they are combined into one thesis. In what follows, I briefly outline each of the articles, before I integrate them into a common narrative thread.

Article 1

In the first article, Thor Grünbaum and I begin from a common assumption, namely that some kind of control is necessary for moral responsibility. For that assumption to be justified, humans should be psychologically capable of realising this kind of control. We provide a positive account which integrates philosophical theories of moral responsibility and action with a cognitive psychological explanation of the mechanisms that realise the capacity for cognitive control. This article is a contribution to the growing literature which brings philosophy of free will and moral responsibility, philosophy of action, and philosophy of psychology together with the cognitive science of control (see Shepherd, 2014, 2015, 2016; Buehler, 2019, 2021; Murray & Vargas, 2020).

The article begins by spelling out the relevant kind of control that is necessary for moral responsibility. We call this *responsibility-level control*, in contrast to other notions of control used in philosophy and the cognitive sciences. One way to determine whether an agent has responsibility-level control is to track her agentive contribution to an action. Thus, a good explanation of responsibility-level control should leave space for a contribution by the agent. However, the nature of psychological explanations motivates a form of scepticism about the psychological plausibility of responsibility-level control. Typically, psychological explanations by

functional decomposition explain a capacity by breaking it down into smaller parts, often low-level mechanisms. Such explanations leave no space for a significant contribution by the agent because the agent cannot be identified with any of these mechanisms. The absence from the psychological explanation of a contribution by the agent would suggest that responsibility-level control is not within the range of human psychology. Such a prospect invites scepticism about whether people are morally responsible in the first place (Levy, 2011: 103-106, Levy, 2014: 109-126, Pereboom, 2006: 120–22).

Finally, we offer a reply to the challenge. Integrating prominent psychological models of cognitive control with recent philosophical interpretations of such models, we propose an account of standing intentions as structuring causes of action. We locate the agent's contribution in certain higher-order mental states which play the role of biasing attention and determining values of cognitive control. Our proposal allows for the agent to have a significant contribution in bringing something about.

Our explanation of responsibility-level control in terms of standing intentions as structuring causes of action strikes a reflective equilibrium between our best psychological evidence on the one hand, and prominent philosophical theories on free will and moral responsibility on the other. Furthermore, this article makes a first step in turning the focus on to the agent. In our view, the relevant sort of control is associated with a contribution by the agent. Thus conceived, control is inextricably agent related. The article is a first step towards arguing that whatever the conditions for moral responsibility end up being, they are inseparable from the psychological and epistemic constraints of the agent.

Article 2

Article 2 shifts the focus to the problem of luck. Luck has long been a thorny issue for theories of free will, moral responsibility, and most accounts of action. Historically, there is a common view that moral practice should be immune to luck (see Kant, 1784 [1998]; Nagel, 1979; Williams, 1981). This view captures the natural idea that good or bad luck should not affect how blameworthy or praiseworthy an agent is. Decisions that are a matter of luck typically do not fulfil the conditions for moral responsibility. Here, I argue that given certain epistemic constraints that ordinary agents face, any decision may be prone to a certain kind of luck which I call *mental luck*. The existence of mental luck has serious implications for prominent theories of moral responsibility.

I begin by presenting a version of the luck problem called present luck (Mele, 2008). While it was meant as a challenge primarily for libertarian theories of free will, I introduce a proposal which generalises present luck beyond libertarianism. This epistemic version of the problem of luck, which I call mental luck, affects all agents regardless of whether the universe is deterministic or indeterministic. Mental luck suggests that because of our limited epistemic perspective of the world, we make decisions under conditions of subjective uncertainty. Thus, it is possible for an agent to make different decisions, while her accessible mental states remain identical. Such mentally lucky decisions are not determined by the agent's accessible practical attitudes.

Finally, I consider how mental luck affects theories of moral responsibility. Most theories associate the agent's moral responsibility with her practical attitudes and with the outcome of her decisions. Mentally lucky decisions do not fulfil this condition. Therefore, decisions under the influence of mental luck are not free and morally responsible. Importantly though, because of our limited epistemic perspective, any decision may be prone to mental luck. This conclusion poses a serious challenge to prominent theories of moral responsibility, and points to a dilemma about the centrality of conscious awareness for the purposes of moral assessment.

Article 2 builds on the argument made in the first one, namely that the contribution of the agent is an integral part of the sort of control that is necessary for moral responsibility. Having turned the focus on the capacities, constraints, and perspective of the agent, we are now faced with the problem of luck. Regardless of the objective nature of the universe, we all act and decide in conditions of relative uncertainty. One way to understand the implications of this predicament is through the worry about mental luck.

Article 3

In the final paper that forms part of this dissertation, I defend the claim that people can be morally responsible for actions and their consequences. To do that, I turn again to the issue of luck, more specifically resultant moral luck. Common cases of resultant moral luck apparently demonstrate that luck can affect the outcome of an action. In such cases, an agent may be less or more blameworthy because of factors outside her control. Thus, resultant moral luck poses a challenge to most theories of moral responsibility, which maintain that factors beyond the agents' control -such as luck- should not affect the moral evaluation of agents.

Cases of resultant moral luck push us into an unattractive dilemma between internalism (i.e., Khoury, 2018), which would suggest that people are only morally responsible for their mental willings, and scepticism (Levy, 2011). However, I argue that the force of resultant moral luck cases rests upon a rarely explicit assumption, namely that actions are events. I spell out this event ontology of action and consider why it is not the best way to capture ordinary actions. I move on to discuss an alternative ontology of action, which departs from the structure of *one-shot* events towards describing action as ongoing goal-directed process with a *many-shots* structure. Described under the alternative ontology, prominent cases of resultant moral luck appear as special cases, and thus no longer representative of ordinary action. The force of the challenge by resultant moral luck is diminished, and so is the credibility it provides to the view that people are not morally responsible for their actions.

In the final part of article 3, I turn to predictive coding as a plausible, naturalistic proposal for human cognition. I argue that predictive coding is consistent with the *many-shots* structure of action. Within a cognitive framework primarily engaged in error minimisation, errors are not necessarily failures, but rather ordinary parts of the process of action. Predictive coding provides a realistic proposal that is consistent with the process ontology of action. Thus, it provides support to the claim that cases of resultant moral luck do not show that people are not morally responsible for their actions or their consequences.

While it may seem counterintuitive at first, this last article presents the positive account of the dissertation. With the capacities and limitations of the psychological agent in given the primacy they deserve; we realise that we are limited by our perspectival view of the world. For instance, we are bound by luck, and this challenges our notion of moral responsibility. However, there is reason to think that ordinary action approximates an ongoing, many-shots process of constant error minimisation. And this seems to be consistent with a cognitive framework whose very purpose is to manage errors, luck, and uncertainty.

A Common Narrative

One of the core components of my work is the idea that agents operate within an environment of limited information and subjective uncertainty. This predicament turns decision-making and action into a game of probabilities. With limited information at hand, and a limited capacity for control, one is significantly restricted. Traditionally, we have been thinking of action as deciding, then doing. Perhaps a more realistic approximation is guessing, then trying, then trying again. There are important issues with moral responsibility in a probabilistic setting, even if it is probabilistic only from the perspective of the agent. In fact, this is the overarching theme of my dissertation: How is the type of control necessary for moral responsibility possible in a probabilistic setting of subjective uncertainty and limited information? The first step in answering that is establishing the importance of the agent contributing to an action, which is the focus of the first article.

Next, one needs to establish that the answer will not depend on whether the universe is deterministic or not. Because of epistemic and psychological limitations, our decisions involve a degree of subjective uncertainty, such that the truth of determinism does not matter. That is the starting point of the second paper. What does this predicament mean for the agent? I arrive at the worry about mental luck, which suggests that agents may end up in different decision outcomes, while having the same reflectively available mental states. From the limited perspective of the agent, even the inner sanctum of decision-making is prone to some form of luck. The worry poses a challenge for most theories of moral responsibility. Does that mean that moral responsibility is impossible? Not necessarily, because it presupposes a specific idea about what actions are.

That is the starting point of the third and final article. Here, I consider another kind of luck which influences the outcome of actions, namely resultant luck. Cases of resultant luck depend on conceiving of actions as events. Instead, I argue that ordinary action is best understood as a *many-shots* process. Given the epistemic outlook of agents in the world, even if the universe is deterministic, it is credible that control would have evolved in such a way that agents can exercise it within a context of uncertainty. One way to understand that is to think of actions as something that progressively grows over time. Within this ontology, human thought and action can fit into a probabilistic cognitive framework, where the primary task of the mind is to minimise errors. The proposal suggests that our cognitive capacities exist to enable us to exercise control over an environment of uncertainty and limited information, and to minimise the influence of luck on our actions.

References

Aristotle (1985). *Nicomachean Ethics* (Irwin, T. trans.). Indianapolis: Hackett

Balaguer, M. (2012). *Free will as an open scientific problem*. MIT Press.

Björnsson, G. (2017). Condition on Moral Responsibility. *Responsibility:* In Wieland, J. W., & Robichaud, P. (Eds.). (2017). *Responsibility: The epistemic condition*. Oxford University Press.

Buehler, D. (2019). Flexible occurrent control. *Philosophical Studies, 176*(8), 2119-2137.

Buehler, D. (2021). Agential capacities: a capacity to guide. *Philosophical Studies*, 1-27.

Chisholm, R. (1966). *Might We Have No Choice? In Keith Lehrer*, (ed.), *Freedom and Determinism*. New York: Random House

Clarke, R. (1993). Toward a credible agent-causal account of free will. *Noûs, 27*(2), 191-203.

Clarke, R. (2003). *Libertarian Accounts of Free Will.* New York: Oxford University Press.

Dennett, D. (1984). Elbow Room: The Varieties of Free. *Will Worth Wanting*. Cambridge, MA: MIT Press

Edwards, P. (1958). *Hard and soft determinism*. In Sydney Hook, ed., *Determinism and Freedom in the Age of Modern Science.* London: Collier Books.

Ekstrom, L. (2000). Free will: A philosophical study. Boulder, CO: Westview

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy, 66*(23), 829–839.

Frankfurt, H. G. (1978). The problem of action. *American philosophical quarterly, 15*(2), 157-162.

Ginet, C. (1980). The conditional analysis of freedom. In *Time and cause* (p. 171-186). Springer, Dordrecht.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Ginet, C. (1996). In defense of the principle of alternative possibilities: Why I don't find Frankfurt's argument convincing. *Philosophical Perspectives*, *10*, 403-417.

Ginet, C. (2000). The epistemic requirements for moral responsibility. *Philosophical Perspectives*, *14*, 267-277.

Ginet, C. (2003). Libertarianism. In *The oxford handbook of metaphysics*.

Haji, I. (2002). C*ompatibilist views of freedom and responsibility.* In Kane, R. (ed.), *The Oxford Handbook of Free Will*. (p.202-229). Oxford University Press.

Haji, I. (2009). *Incompatibilism's Allure: Principle Arguments for Incompatibiism*. Peterborough, Ontario: Broadview Press.

Hartman, R. J. (2017). *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness*. Routledge.

Hobart, R. E. (1934). Free will as involving determination and inconceivable without it. *Mind*, *43*(169), 1-27.

Hume, D. (1955). *An Inquiry Concerning Human Understanding: With a Supplement, an Abstract of a Treatise of Human Nature*. Bobbs-Merrill Educational.

Kane, R. (1995). Acts, Patterns, and Self-Control. *Behavioral and Brain Sciences*, *18*, 131-2.

Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.

Kane, R. (2002). *Introduction: The contours of contemporary free will debates*. In Kane, R. (ed.), *The Oxford Handbook of Free Will*. (p.3-41). Oxford University Press.

Kant, I. (1784) [1998], *Groundwork of the Metaphysics of Morals* (Gregor, M. ed. and trans.). Cambridge: Cambridge University Press.

Kant, I. (1998). *Critique of pure reason* (Guyer, P. & Wood, A.W. ed. and trans.). Cambridge: Cambridge University Press.

Khoury, A. C. (2018). The objects of moral responsibility. *Philosophical Studies*, *175*(6), 1357-1381.

Lamb, J. W. (1977). On a proof of incompatibilism. *The Philosophical Review*, *86*(1), 20-35.

Lang, G. (2021). *Strokes of Luck: A Study in Moral and Political Philosophy*. Oxford University Press.

Latus, A. 2000, Moral and Epistemic Luck, *Journal of Philosophical Research*, 25, 149–172.

Latus, A. (2003). Constitutive luck. *Metaphilosophy*, *34*(4), 460-475.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

McGeer, V. (2012). Co-reactive attitudes and the making of moral community. *Emotions, imagination, and moral reasoning*, 299-326.

McKenna, M., & Pereboom, D. (2016). *Free will: A contemporary introduction*. Routledge.

Mele, A. R. (1995). *Autonomous Agents: From Self Control to Autonomy*. Oxford University Press.

Mele, A. R. (1996). Soft libertarianism and Frankfurt-style scenarios. *Philosophical Topics*, *24*(2), 123-141.

Mele, A. R. (2006). *Free will and Luck*. Oxford University Press.

Mele, A. R. (2010). Moral responsibility for actions: Epistemic and freedom conditions. *Philosophical Explorations*, *13*(2), 101-111.

Moore, G. E. (1912). Free will. *Ethics,* 84-95.

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies*, 177(3), 825-843.

Nagel, T. (1979). *Moral Luck*. In Nagel, T., *Mortal Questions*. Cambridge: Cambridge University Press, 24-38.

O'Connor, T. (1996). Why agent causation?. *Philosophical Topics*, *24*(2), 143-158.

O'Connor, T. (2000). *Persons and causes: The metaphysics of free will*, New York: Oxford University Press.

Pereboom, D. (1995). Determinism al dente. *Noûs*, *29*(1), 21-45.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Posner, M. I., & Snyder, C. R. R. (2004). *Attention and Cognitive Control.* In D. A. Balota & E. J. Marsh (eds.), *Key readings in cognition. Cognitive psychology: Key readings* (p.205–223). Psychology Press.

Article 1:
Agentive Contribution, Control, and Moral Responsibility

Co-authored by Thor Grünbaum

This article is under review at the Journal Philosophical Psychology, after a request for revisions.

Abstract

Can humans exercise psychological control sufficient to realise the kind of control necessary for being morally responsible? The actual psychological control mechanisms involved in thought and action are described by cognitive psychology and neuroscience. By contrast, the control necessary for moral responsibility is the object of common-sense moral psychology and philosophical reflection. This kind of responsibility-level control is challenged by the form of explanation in cognitive psychology: the form of functional-mechanistic explanations leaves no space for agency, as personal-level capacities are explained by sub-personal, non-agentive mechanisms. Since most prominent theories of moral responsibility require that people contribute agentively to their actions in order to be morally responsible, this challenge motivates scepticism about the possibility of moral responsibility. By integrating insights from psychological models of cognitive control and philosophy of action, we provide a way to think of the agent's contribution in functional-mechanistic terms. We argue that a set of executive mechanisms that constitute an agent's cognitive control of thought and action also realise responsibility-level control.

Section 1: Introduction

Control is commonly accepted as a necessary condition for moral responsibility. Are humans the sort of psychological creature that can realise the right kind of control? We attempt to answer this question by integrating philosophical accounts of moral responsibility and theories of action with a cognitive psychological explanation of the mechanisms that realise our capacity for cognitive control. Recent work has sought to bring together theories of moral responsibility, along with issues

in philosophical psychology of action, and the cognitive science of control (see for instance, Shepherd, 2014, 2015, 2016; Buehler, 2019, 2021; Murray & Vargas, 2020). Our paper is a contribution to this literature.

Psychology challenges theories of responsibility in at least two distinct general ways. First, some types of empirical results indicate that humans are not psychologically capable of realising the sort of control that is necessary for moral responsibility. For instance, results from Libet's timing experiments challenge common intuitions about conscious choice (for a discussion of this form of challenge see Schlosser, 2013). Second, the nature of psychological explanations might motivate a form of scepticism about the possibility that humans can psychologically realise a capacity for this sort of control. This challenge flows from the very form of explanation by functional decomposition, rather than any particular set of surprising results. Here, we address the second challenge.

Scepticism about the psychological realisability of control leads to scepticism about moral responsibility in general. The prospect that the right sort of control is not within the range of human psychological capacities invites a degree of doubt about whether we are morally responsible in the first place (Levy, 2011: 103-106, Levy, 2014: 109-126, Pereboom, 2001: 120–22). We aim to alleviate such sceptical worries, at least in part. We provide an explanation of control that is consistent with both our best psychological evidence and also with prominent philosophical theories on free will and moral responsibility.

We begin in Section 2 by describing the type of control that is often considered by both theories of moral responsibility and common sense to be necessary for attributions of blame and praise. We call this *responsibility-level control,* in contrast to other notions of control used in philosophy and the cognitive sciences. We spell out responsibility-level control in terms of agency. Specifically, we propose that at least one way to determine whether an agent has the right kind of control to be morally responsible is to track her agentive contribution to an action. While this is not the only way to understand responsibility-level control, it brings out an important feature of moral responsibility. How much an agent contributes to the bringing about of some event is central to the question of whether the agent has responsibility-level control of the event and, therefore, whether she is morally responsible for bringing about the event.

In Section 3, we consider a challenge to the idea that humans are psychologically capable of realising responsibility-level control. The challenge is motivated by a common understanding of the nature of psychological explanations. According to this understanding, a functional or mechanistic explanation of the capacity for control leaves no space for the agent to have any significant contribution in causing something to happen. And since agentive contribution is partly constitutive of responsibility-level control, its absence from the psychological explanation would suggest that such control is unrealisable by human psychology.

In Section 4, we offer a possible reply to the challenge. By reviewing recent psychological models of cognitive control, as well as recent philosophical interpretations of these models, we propose an account of standing intentions as structuring causes of action. We argue that the agent's contributions can be found in the role of higher-order mental states in biasing attention and setting values for parameters of cognitive control. Our proposal is consistent with the agent having some significant contribution in bringing something about. We finish the paper by discussing a number of potential objections, before we wrap things up in the conclusion.

Section 2: Responsibility-level Control and Agentive Contribution

If attributions of moral responsibility require responsibility-level control and we are morally responsible, we should be able to psychologically realise this sort of control. This is a central tenet of the paper. But before we can turn directly to the question of whether humans are the type of creatures that can psychologically realise such control, we need to explicate responsibility-level control further.

Responsibility-level control is often equated with the freedom condition (for instance in Mele, 2010). It is whatever kind of freedom is necessary for being morally responsible. Typically, an agent has responsibility-level control over X only if she can exercise certain kinds of freedoms over X. For instance, she may have the freedom to choose X, or the freedom to inhibit doing X, or the freedom to make X happen. Responsibility-level control refers to freedoms relevant for moral responsibility. In this section, we look into one aspect of responsibility-level control which we take to be independent of one's metaphysical commitments concerning freedom. We focus on agency, and more specifically on the importance of the agent's contribution to an action for the purposes of moral assessment.

Agency features in discussions of moral responsibility because it concerns the degree to which an agent is doing something. For the rest of the paper, we refer to the relevance of agency for the purposes of moral assessment as her *agentive contribution*. The central claim of this section is that the degree of an agent's agentive contribution to the occurrence of an event or action is crucial in determining the degree to which the agent is morally responsible for her action. When the agent's contribution with respect to the occurrence of some action or event approximates zero, so does the degree to which she is morally responsible for said action or event. First, the section argues that at least one central form of attribution of moral responsibility targets what we are calling the agent's agentive contribution. Second, that the agent's degree of moral responsibility for her action is related to her degree of agentive contribution.

One form of ascription of blame or praise concerns actions performed by an agent. The objects of this form of moral attribution are typically events an agent is actively bringing about. If such an event were the product of total coincidence, bad luck, or external force, the agent is no longer blameworthy or praiseworthy for its occurrence. Intuitively, people are responsible for what they do, not for what is merely happening to them. The degree to which an agent is doing something is her *agentive contribution*.

One way to understand agentive contribution is in terms of basic action. Consider the example of an agent intentionally flipping a light switch to alert her secret lover. What exactly is the agent's basic action in this case? Different positions diverge on this point. First, some philosophers of action equate basic action with bodily movement. Davidson's claim that "we never do more than move our bodies: the rest is up to nature" (Davidson, 2001: 59) is indicative of this position. Here the agent's contribution lies in her moving her body intentionally, that is, moving her finger. Second, some theories individuate basic action with reference to the agent's practical reasoning and skills (Enç, 2003: Ch. 2). According to this type of theory, what an agent basically does is whatever she can do without having to do something else intentionally in order to do it. Here the agent's contribution might be her instrumental object-involving action. She flips the switch as a basic action (Grünbaum, 2013). Third, some theories view agentive contribution as almost unrestricted. According to this view, the notion of basic action is incoherent (Lavin, 2013). Alerting her secret lover could be the agent's genuine contribution (Ford, 2018). Fourth, some theories identify the agent's real contribution with her guaranteed and failure-proof agency. According to this final type

of theory, all the agent can do is mentally try or will the action, and "the rest is up to nature" (Hornsby, 1980; for a discussion of volitionalism, see Grünbaum, 2008).

It is not important to our argument which of these accounts correctly locates the agent's contribution to the bringing about of some event. For us, the important aspect is that when locating agentive contribution theorists are concerned with how much luck to tolerate. Assuming a causal account of action, an event is an agent's action if it is caused in the right way by the right constellation of mental states. If the event is brought about by external forces or by luck, it does not satisfy that condition (Davidson, 2001: Essay 4; Mayr, 2015: Ch. 5). The further into the world we extend the agentive contribution, the less active guidance and control seem to be involved and the more vulnerable the process becomes to external forces and chance.

One does not need to accept a causal theory of action to be swayed by these considerations. Merely accepting that to act requires knowledge that one is acting might lead to adopting similar constraints. Locating the agent's contribution depends on what the agent can know (Grünbaum, 2009; Mele & Moser, 1994). The more agentive contribution extends beyond the agent, the less plausible it seems that she knows what she is doing. Considerations like these have persuaded some theorists to move the extent of agency inwards in various steps, towards the mental life of the agent.

One's theoretical understanding of agentive contribution and where it is located affects what one accepts as the legitimate object of moral appraisal. Concerns about failure and breakdown that move some theorists to restrict the scope of the agent's agentive contribution to her bodily movements or her mental volitions are the same concerns that might move one to restrict the object of moral appraisal. Because these breakdowns are beyond the agent's control, it seems natural to think that consequences which lie beyond the agent's real contribution are prone to moral luck. If one rejects moral luck, one might limit the object of moral appraisal to the most minimal and failure-proof contribution of the agent (Khoury, 2018).

One can think of the whole spectrum of cases, from full-blooded agency to unreflective, habitual action as a matter of degrees of agentive contribution. Any decrease in agentive contribution corresponds to a decrease in responsibility-level control and moral responsibility. One common way to illustrate this point is via excusing conditions, namely factors that diminish the extent to which an agent is morally responsible. These are factors that typically attenuate the blameworthiness of an agent.

Typically, circumstances that may count as excusing conditions include performing an action as a result of being forced, manipulated, or unconscious. Imagine a person that ruminates for days in making a decision which issues in a simple but momentous act of signing her name (O'Shaughnessy, 2000: 101). As a fully reflective, full-blooded action where the agent is carefully considering all the morally and prudentially relevant reasons for signing her name, the agent is fully responsible for putting her name to the paper and for its morally relevant consequences. Now imagine variations on the case, where the agent signs her name because she has been tricked into having a number of false beliefs, because of external coercion (at gun point), because she has been hypnotised, or because another agent is stimulating her neural circuits. All are cases of diminished agentive contribution, where the agent is less than fully blameworthy or praiseworthy.

Cases like these drive a core intuition about being morally responsible, namely that it must involve some degree of contribution by the agent. When a person signs her name because she has been hypnotised, the agent's contribution was merely allowing herself to be hypnotised by the malicious hypnotist. Granted, it was she who signed her name, but while under the hypnosis, we might think that her contribution to the movements of her body is reduced to that of an automaton, moving in all the ways for which the designer planned. It is precisely this lack of agentive contribution to the bodily act of signing that serves as an excusing condition. The less of a role an agent has in bringing about an action, the less morally responsible she seems to be.

Now to a potential objection. Do cases like these indicate that what we call *agentive contribution* is simply what philosophers of free will call *ability to do otherwise*? This depends on one's metaphysical interpretation of the agent's contribution. For some free will libertarians, the agent's contribution should be understood in terms of an ability to do otherwise, an ability which is incompatible with determinism (see for instance Steward, 2012). However, we intend the notion of agentive contribution to be independent of these metaphysical discussions. Irrespective of one's position on freedom of the will, moral responsibility concerns the actions of an agent.

The importance of excusing conditions for understanding the role of agentive contribution in moral assessment is further corroborated by the discussion on the difference between a willing and an unwilling addict (Frankfurt, 1971). Addict cases illustrate why responsibility-level control is connected to agentive contribution and why the latter may come in degrees.

Suppose two people are addicted to the same drug X. Both addicts have a first-order desire to take the drug, a direct implication of being addicted. Furthermore, they both habitually use X. Lastly, both have the capacity to engage in second-order thoughts about their use of X. So far, the two addicts are psychologically indistinguishable. But as Frankfurt (1971) has suggested, we identify with some beliefs and desires more than others. The willing addict has a second-order desire which endorses her first-order desire to take X. She endorses her addiction, while the unwilling addict repudiates it. Her second-order motives do not align with, but rather contradict her first-order desire to take X.

Does the difference in second-order motivation matter for how we assign moral blame to the two addicts? Frankfurt and others (for example Sripada, 2017) think it does. Unwilling addicts may be at least partly excused, because their second-order motives to refrain from their addiction are decoupled from their first-order desire to indulge. They are unable to align their behaviour with their second-order motives. But that second-order motive better reflects their true self, while the first-order desire is more in some way more alien to them.

To be sure, not everyone accepts Frankfurt's hierarchy of motives as the correct marker of agency. As several philosophers have pointed out (for example Velleman, 1992; Watson, 1975) there is no reason to think that second-order motives cannot be alien to the agent, just like any other motive or desire. For our purposes, the outcome of this debate is not important. The two addicts illustrate a difference in the degree to which each of them contributes to an identical action. Endorsing her addiction, the willing addict is in some meaningful way contributing significantly more to her action than the unwilling one is. As a result of her higher degree of contribution to her actions, she seems to be more morally responsible. The more an agent contributes to her addiction (in this case by endorsing it), the more blameworthy one is. As agentive contribution decreases, so presumably does control, and one is more likely to be excused.

Before closing this section, let us mention a potential worry relating to agentive contribution as a yardstick for measuring the degree of responsibility-level control. We propose that degree of agentive contribution is related to degree of control, and, therefore, necessary for moral responsibility. But what if there are cases where agentive contribution and moral responsibility come apart? For instance, one could appeal to common examples in the moral responsibility literature, such as omissions, forgetting to perform an action, or various unconscious attitudes. Such cases seem to indicate that an agent can be morally assessed for something to which she has not

contributed agentively. Does that provide reason to think that agentive contribution and moral responsibility come apart? We think it does not. Our proposal is that the degree to which an agent contributes to something happening co-varies with her responsibility-level control. Cases such as the ones above are indeed challenges for control theories of moral responsibility, but for our purposes, the solution to these challenges is not important. Whatever strategy the control theorist would adopt to defend her position, that strategy would also be available to us. To the extent that a control-theorist can successfully respond by distinguishing between direct and indirect responsibility (Zimmermann, 1997; Nelkin & Rickless 2017; Mele 2020), or by grounding moral responsibility for omissions on agential capacities (for instance Clarke, 2014), we can avail ourselves to these strategies too.

Section 3: A Challenge to Agency

So far, we have argued that the sort of control that is necessary moral responsibility is related to the degree to which an agent is contributing to an action. Tracking the one's contribution is one way to determine whether one has responsibility-level control. Relating responsibility-level control to this aspect of agency in our philosophical psychology of agency can help us address the question whether humans have the psychological capacity to realise responsibility-level control. If what we have argued so far is true, then humans realise responsibility-level control and can be morally responsible for their actions only if their actual psychology is such that they can make a real agentive contribution to their actions.

Any attempt to provide a psychological explanation of agentive contribution is faced with a challenge. Central to this challenge is a worry about the kind of explanations that psychology can provide. In its simplest form, the challenge suggests that once we have described the mechanisms that explain our psychological capacity for control, we have effectively explained away the agent's contribution. In other words, it is part of the nature of any mechanistic psychological explanation to leave no space for agentive contribution. And since the latter is central to responsibility-level control, we end up at an uncomfortable position: our best explanations of how control is realised by human psychology cannot accommodate the sort of control that is commonly considered necessary for moral responsibility.

First, let us position the challenge we raise here in the literature. It is more limited in scope than familiar exclusion arguments, which challenge non-reductive physicalism by calling into question the causal power of irreducible mental states (Kim, 1998). Whereas the exclusion argument targets a metaphysical issue about mental causality, we are interested in the relationship between common intuitions (philosophical or otherwise) about psychological phenomena and scientific psychological accounts that explain the same phenomena. In contrast to Kim's argument, the argument we have in mind does not assume that all physical effects have sufficient physical causes, or the completeness of physics. We are focusing on the more specific problem of integrating our common-sense psychological explanations of human action with scientific explanations of human action at various levels of an explanatory hierarchy (Bermúdez, 2004). The problem arises when the functional descriptions of common-sense psychology and scientific psychology do not align. Once we have completed the functional decomposition of the psychological capacities of interest, it might turn out that no components at the lower levels of the mechanism correspond to functional roles identified by common-sense psychology (Bermúdez, 2006; Stich, 1978). Motor control has often been the target of this kind of strategy. Once we descend the decompositional mechanistic hierarchy, intentions, practical reasoning, and rational control seem to disappear (Spencer, 2007; Uithol, Burnston, & Haselager, 2014; for discussion, see Grünbaum, 2012).

In its simplest form, the issue is this: our theories about the requirements of free and morally responsible action should interface with our best psychological accounts of action. If moral responsibility requires that the agent has a certain type of control over an action (one that we have specified in terms of making an agentive contribution), then that type of control should fit within a psychological account of action. If there is a deep conflict between the two forms of explanation, we should plausibly bet our money on the scientific form of explanation.

In motivating this challenge, we focus on psychological explanations that are mechanistic in nature. By mechanistic explanation we refer to a standard form of computational or functional explanation in cognitive psychology and cognitive neuroscience (Cummins, 1983: Ch. 1). Scientific psychological explanations are mechanistic in the sense that they employ functional decomposition to explain some human trait, such as the capacity for cognitive control (Piccinini & Craver, 2011).

Consider a future complete scientific psychological explanation of human reasoning, decision-making and action. As Daniel Dennett and others have pointed out, any explanation of intelligent behaviour should refrain from explaining said behaviour in terms of intentionality or intelligence

(Dennett, 1975). Instead, we should opt for functional-mechanistic explanations, in which higher-order phenomena are explained in terms of lower-order entities and their activities. In scientific psychological explanations, higher-level behaviour is to be explained in terms of a lower-level mechanism (Craver, 2007). The entities and their activities in this mechanism should be explained by mechanisms at a yet lower level. At some level of explanation, descriptions of the entities and their activities would not employ psychological terms.

This is exactly where the challenge appears. Suppose we want to develop a psychological explanation for some intelligent behaviour that humans demonstrate, say their competence in the Wisconsin Card-Sorting Test (WCST) (Grant & Berg, 1948). WCST is a neuropsychological test that measures people's ability to learn and act flexibly, in the face of changing patterns of reinforcement (Monchi et al., 2001). Participants are presented with cards, which must match without knowing the rules. What they do know is whether their last matching attempt was right or wrong. An explanation of what makes people good at the WCST cannot begin by postulating some "problem-solving module". Rather, it would proceed by breaking the task into its constitutive cognitive sub-tasks. Perhaps one of them would be a visual discrimination module that identifies different card types. Another could be a short-term memory module that stores and retrieves information about which matchings are right and which are wrong.

One can suppose many such modules. But with a closer look, it becomes hard to refer back to the agent and her contribution, i.e., the person who is trying to get the matchings right every time. When providing a mechanistic explanation, we are looking for what kind of computation takes place at the lowest level we can explain functionally. As a result, we explain a person's agentive contribution by suggesting that it is computed by a sub-personal mechanism.

If we explain behaviour by appeal to mechanisms found on the sub-personal level, how can we distinguish between what an agent does, and what merely happens to her? There may be certain functional differences between the different levels of mechanisms, but the lower we go, the less likely we are to identify whatever is going on with something the agent is doing. Somewhere in the process of providing an explanation we seem to lose the connection between the agent's personal-level states and agentive contribution, and the mechanisms that realise the behaviour we try to explain. If whatever enables and causes people to make the correct matchings in the WCST takes place at the sub-personal level, then successful matching of two cards no longer seems to be something the agent is bringing about, because there is no space left for the agent to have a personal

contribution. Matchings are performed by people, and yet we end up with a form of explanation according to which matchings happen because of the interplay between a suite of components, of which the agent might have no knowledge.

Something inherent in this type of mechanistic psychological explanation seems to diminish the role of the agent and her agentive contribution. So, as the explanatory project is being realised, one might be less and less likely to identify the agent's contribution with any part of the mechanism that explains an action. Since most work in cognitive psychology follows such a functional-mechanistic approach, we could end up with an explanation where the agent and her contribution are absent. In a way, the issue is simple: As we provide a scientific explanation by accounting for mechanisms further and further down, the agent is less and less part of that explanation.

All this can be summarised in the following argument. Call this the *Argument from Mechanistic Explanation*:

1. If mechanistic psychological explanations are true, personal agency is computed by non-agentive mechanisms.
2. If personal agency is computed by non-agentive mechanisms, then a person's agentive contribution is explained by something other than an agent.
3. Therefore, if mechanistic psychological explanations are true, then a person's agentive contribution is explained by something other than an agent.

If the conclusion is true, the agent is not making any real contribution. Her actions are explained by non-agentive computational mechanisms. As we have seen, there are good reasons to accept premise (1), namely that mechanistic explanations spell out higher-order personal choices and actions in terms of lower-level, sub-personal, non-agentive mechanisms. When it comes to an agent's contribution in bringing something about, a mechanistic psychological explanation will spell it out in sub-personal, non-agentive terms.

Such an explanation will not satisfy those who think that being morally responsible requires a sort of control that allows for an agent to have a personal-level contribution. Responsibility-level control requires that something the agent is doing is making a difference. However, when we move down several levels, it becomes hard to see the agent doing anything at all. Therefore, the argument from mechanistic explanation suggests that responsibility-level control is unrealisable by human psychology.

One way to resist this claim is to show that there is no real conflict between a common-sense explanation according to which the agent is making an agentive contribution and a scientific form of mechanistic explanation. Responsibility-level control is related to the degree of the agent's contribution to an action. At the same time, our best account of human psychology is mechanistic. Therefore, our mechanistic explanation of control should allow for the agent to make some degree of agentive contribution. In order to do that, one option is to deny premise (2) of the argument from mechanistic explanation, namely, that a person's agentive contribution is explained by something other than an agent. Such an explanation should satisfy the requirements of both mechanistic psychological explanations and responsibility-level control. In Section 4 we provide the outline of such an account.

Section 4: Cognitive Control

We have discussed a challenge to the view that responsibility-level control is realised by actual human psychology. The challenge is captured by the argument from mechanistic explanation, which suggests that any mechanistic explanation of the psychological mechanisms that are supposed to realise responsibility-level control is inconsistent with the agent making a real agentive contribution. What is identified as agentive contribution by our common-sense psychological explanation is the product of multiple levels of psychological mechanisms, and the agent cannot be identified with any of them. Our aim in this section is to demonstrate that contemporary cognitive science of control provides the means of avoiding the conclusion of the argument from mechanistic explanation. Even when we focus on lower levels of computational explanation there is still an important role for the agent. In fact, a plausible contemporary scientific contender for a psychological explanation of cognitive control is consistent with the agent making a real contribution.

Our strategy is as follows. We begin by spelling out the agent's contribution in terms of structuring causes of action. Then, we present a prominent scientific psychological model of control, which provides the foundation of our account. With the psychological account laid out, we provide a philosophical interpretation of the model that allows the agent's standing intentions to function as causes of action. We elaborate on this interpretation further and argue why it allows the agent to make an agentive contribution, while still providing a mechanistic explanation of control.

Finally, we consider how our account fares against the argument from mechanistic explanation and how it aligns with common intuitions about moral responsibility.

Two clarificatory remarks are needed. First, note that the challenge to moral responsibility raised by the argument from mechanistic explanation rests on the assumption that the account of responsibility-level control is inconsistent with the psychological explanation of control (since on the best psychological account, we find no real agentive contribution). A plausible reply does not require that we fully spell out the mechanisms by which the agent plays a substantial role in action. Rather, we need to show only that a substantial role for agency is consistent with the psychological mechanisms that enable responsibility-level control. Second, the account needs to explain how the agent might play a role, and where in the process that role would be located. This last part is important because for any account to count as an alternative, it needs to do so in functional-mechanistic terms. Unless we can point to specific mechanisms that might enable the agent's contribution and specific points in the process where that happens, the task is not complete.

Having set the stage, we can now more directly address the issue of the causal relationship between the agent's mental life and her ability to contribute to and guide her actions. Consider Dretske's distinction between two kinds of causes, *triggering* and *structuring* (Dretske, 1989). This is perhaps best illustrated by the example of a thermostat. When temperature drops below some threshold, the thermostat turns on the heating. The drop of temperature is a triggering cause for the thermostat to turn on. The event of a drop in temperature causes the turning on of the heating. However, the thermostat can only do this because someone in the past arranged it in a specific way. So, the structuring cause for the thermostat to turn the heating on at a specific temperature is the way it is designed and internally organised, not the drop in temperature itself.

While both types of causes are usually involved in action, structuring causes are of particular interest to our argument because they refer to the way the system is organised. In the thermostat case, the structuring cause refers to the physical set-up of the device. But in the case of human behaviour, that can be the mental (cognitive and conative) set-up of the agent. Mental organisation can provide structuring causes that regulate behaviour in ways that align with the agent's other motives. In that sense, these mental structuring causes are akin to notions like *future-directed intentions* and *policies* (Bratman 1984, 1987: Pacherie, 2006, 2008), *second-order desires* (Frankfurt, 1971, for a discussion see Mele, 1992), or *values (*Watson, 1996). Rather than directly

triggering mental events, these standing intentions, policies, desires, and values refer to continuing long-term states of the agent.

An agent's contribution is more than just the triggering causes of her actions. Often, we think of an agent's doing something in terms of mental events causing the action initiation (i.e., as triggering causes of action). If one thinks of such contribution solely in terms of triggering causes, it might be difficult to counter the argument from mechanistic explanation. But there is a different way to think of the agent's contribution. She manages to steer her life in a particular direction. She shapes and configures her actions and reasoning according to her goals, values, intentions, policies. As soon as one thinks about agency it this way, the types of causes that should be of interest are more akin to structuring causes of action.

In cognitive psychology, control is studied in contrast to behaviour that is automatic or habitual. That is, cognitive control often refers to our ability to "pursue goal-directed behaviour, in the face of otherwise more habitual or immediately compelling behaviours" (Cohen, 2017: 3). For most experimental paradigms which study cognitive control and for the models that accompany them, the distinction between controlled and automatic processing is fundamental. According to a standard conception, controlled processes are generally slower and subject to interference by automatic processes (Shiffrrin & Schneider, 1977; Posner & Snyder, 2004). When a controlled process is obstructed by an automatic one, the two processes compete for limited cognitive resources. As a result, the controlled process is impaired.

Controlled processes being impaired by automatic ones is a well-known effect, famously illustrated by the Stroop task (Stroop, 1935). In the Stroop task, participants are presented with series of colour words. The ink of the words sometimes matches the colour ("red" in red ink) and sometimes it does not ("red" in blue ink). When ink and word colour match, the stimulus is congruent. When they don't the stimulus is incongruent. The Stroop task reliably produces a few results. First, regardless of whether the stimulus is congruent or incongruent, people are generally faster at reading the words than naming the ink colours. Second, incongruent stimuli significantly interfere with people's responses when they are asked to name the ink colour, but not when they asked to read the word. Third, when a secondary task (say, completing arithmetic calculations) is added to the original Stroop task, the primary task suffers. The Stroop interference effect is widely taken as evidence that reading is an automatic process, while naming colours is a controlled one (for a review, see Bugg, Jacoby & Toth, 2008).

A plausible explanation of the interference effect observed in the Stroop task is the *multiple resources hypothesis* (Meyer & Kieras, 1997; Botvinick, Braver, Carter, Barch & Cohen, 2001), which suggests that interference is the reason that cognitive control exists in the first place. As Jonathan Cohen puts it, "a fundamental function of control is to reduce interference where it can arise" (2017: 6). In order to avoid the impairment caused by interference, the control system needs to identify potential cases of conflict between different processes and allocate resources to manage the conflict and interference accordingly. For that, something needs to distinguish between processes that should be granted some of the limited cognitive resources and processes that might interfere with the controlled task and cause an impaired performance.

Consider the case of driving a car. As the driver, you are exposed to various stimuli. However, to safely arrive at your destination, you need to be able to distinguish between potentially relevant stimuli (stop signs, break lights in the distance) and potentially irrelevant stimuli (a seagull flying above), before you can allocate resources in such a way that driving is not negatively affected. According to the multiple resources hypothesis, it is the very purpose of cognitive control to apply constraints on how multiple ongoing tasks are executed, thus avoiding interference where it matters.

Cognitive resources are limited. Allocating control on one process leaves less for others. Consequently, controlled processing comes at an opportunity cost to other processes. One way to understand the relation between cost and opportunity is in terms of effort and motivation (Botvinick & Braver, 2015; Braver 2015). In the Stroop task, there seems to be some alignment between the cognitive control involved in cases of task interference and the degree of effort an agent needs to exercise to minimise that interference. Intuitively, it feels relatively effortless to read "red" in red ink. But naming the colours of incongruent stimuli feels effortful. Such effort indicates the cost involved in reducing interference. Higher effort implies higher cost. Motivation could relate to the opportunity cost of controlled processing by signifying the willingness of the agent to pay that cost (Cohen, 2017). The higher the motivation to engage in a task, the higher the willingness to assume the cost. Effort and motivation are not identical to cognitive control, rather, they are associated with the decision to spend cognitive control resources in a particular way (Westbrook & Braver, 2015).

Models that focus on effort and motivation have recently made significant progress in explaining the mechanisms that underlie cognitive control (see for example Kurzban, Duckworth, Kable & Myers, 2013; Braver, 2015). Such models claim that cognitive control resources are allocated based on evaluations of the expected values of different tasks (Shenhav, Botvinick &

Cohen, 2013). Here resources are allocated in terms of economic decision-making. Different tasks are assigned different expected values, according to their corresponding goals, as well as other criteria. For example, if a task corresponds to a highly valued goal but its likelihood is low, the expected utility will be lower than if the likelihood was high. These assigned expected values (and corresponding calculations of expected utility) determine which goals should be selected, which in turn determines the tasks on which cognitive control resources are allocated (considering the required cognitive effort). According to an influential model proposed by Miller and Cohen (2001), the expected utility of each task is maintained in working memory. From there, and according to the pre-defined values, the selection of processing pathways is biased towards the ones that aligned stimuli with high-valued goals or tasks. In other words, cognitive resources are allocated to behaviours that are more likely to maximise value for the agent.

According to a standard conception in cognitive psychology, cognitive control is the ability to "regulate, coordinate, and sequence thoughts and actions in accordance with internally maintained behavioural goals" (Braver, 2012: 106). This provides a way to specify how cognitive control functions in accordance with internally maintained behavioural goals. Standing mental states and the way they are organised constitutes the agent's short-term and long-term goals, which determine the values decisive for resource allocation in the agent's cognitive control of ongoing tasks.

Now with the model laid out, we need a way to understand what may count as a goal and how the agent is involved in setting, maintaining, and bringing about their goals. A useful starting point can be found in Wayne Wu's account of *agentive control* (Wu, 2016). Here, the agent's standing intentions play a crucial role in the cognitive processing that involves mechanisms at a sub-personal level. Wu sees the agent's mental features as a way to influence how control is applied to some tasks, and not to others. Such personal-level features bias how attention is allocated to specific stimuli, in accordance with their expected utility calculation. Thus, according to Wu, attention functions as the means by which certain stimulus-response associations are selected over others.

As an illustration, consider the following case from a football match. Peter is a skilled goalkeeper, who can save powerful shots coming at him from many directions. He is confronted with a plethora of visual and other stimuli, which he can face with a plethora of motor responses. In order to raise his chances of making a save, Peter needs to mediate his attention so that he attends to the football coming at him. Furthermore, he needs to choose between all the possible responses available. Should he try to catch or to deflect the ball? Peter's attention is biased towards the ball, because the value of the ball is high, in line with his standing intention of winning the game and the

high value he places on being a good goalkeeper. Then, with his attention on the speed and direction of the ball, Peter needs to couple it with an appropriate response, say a leap to the higher-right corner of the goal, in contrast to all the other actions he could perform. Peter's standing intention and higher-order values bias attention to a set of possible stimuli and a set possible motor responses. Attention is on this account a way to solve the problem of coupling many possible stimuli with many possible responses by biasing the selection of a set of couplings.

According to the model of control we sketched above, controlled processing functions by biasing processing pathways towards some (and not other) stimuli. The biasing is the result of top-down signals, which in turn are determined by expected values for each task (which in turn are computed by considering the importance of the goal, the probability of realising it by performing this task, and the intrinsic cost of performing it). The values themselves represent the degree of alignment between a task and a goal. But the goal and its importance to the agent is given by the agent's overall beliefs, values, desires, or intentions. Coming back to Wu's characterisation of the relationship between intentions and attention, we can think of standing intentions as structuring causes of attention. That is, they affect the degree and kind of control that the agent exercises, by structurally causing attention to be allocated to specific objects by biasing processing pathways that couple attentional selection of stimuli with attentional selection of responses (2016: 113).

Now we can depart from and extend Wu's account. The structuring causes involved in setting up the control parameters decisive for the cognitive control of ongoing tasks are not simply mental states such as the agent's standing intentions, desires, and beliefs. We can, thus, begin to understand structuring causes of action as making up the agent's constitution and providing the basis for her agential contribution. This is not the place to engage with the large and complex literature on self-governance and autonomy, but if theories about wholehearted and temporally extended agency along the lines of Frankfurt's higher-order volitions (Frankfurt, 1988: ch.12) or Bratman's self-governance (Bratman, 2004) are correct, we have a way to understand how the agent contributes to an action, by determining the values of the control parameters decisive for resource allocation of cognitive control – that is, decisive for control of thought and action. One way to approach this is to ask what determines the subjective expected values that are assigned to the different tasks where control is expended. Our proposal is that we think of standing future-directed intentions, policies, and higher-order motivation and values as structuring causes that play an important role in biasing or setting the values involved in utility calculations central to resource allocation of cognitive control. For instance, if proposals like the *expected value of control* model (Shenhav et al., 2013)

are correct, these value computations reach all the way down to the level of biasing the neural activation of control in relevant brain areas.

To be sure, this is just a sketch of a possible account. Importantly, open issues remain with respect to how exactly structural causes such as standing mental states and their organisation can flexibly set and adjust the values and biases of cognitive control, how such standing states become occurrent and active in working memory, and what happens in the face of possible conflicting standing intentions (although some of these issues have been addressed in Grünbaum & Kyllingsbæk, 2020; Grünbaum, Oren, & Kyllingsbæk, 2021). The aim here has been more limited, namely to show that a plausible account of personal self-governance, and temporally extended agency is consistent with a prominent computational account of cognitive control.

We have focused on overcoming the challenge raised by the argument from mechanistic explanation by denying premise (2). If we accept the multiple resources hypothesis of cognitive control and our account of standing intentions, higher-order volitions, and values as structuring causes, we have an account that is both consistent with the explanatory principles of a mechanistic cognitive psychology and the claim that agents can make real agentive contributions. This package should be consistent with premise (1) and inconsistent with premise (2).

Let us revisit the *Argument from Mechanistic Explanation*:

1. If mechanistic psychological explanations are true, personal agency is computed by non-agentive mechanisms.
2. If personal agency is computed by non-agentive mechanisms, then a person's agentive contribution is explained by something other than an agent.
3. Therefore, if mechanistic psychological explanations are true, then a person's agentive contribution is explained by something other than an agent.

We should now be able to defend (1) and deny premise (2). Regarding premise (1), it is clearly the case: higher-order agency is explained in terms of lower order mechanisms in the cognitive control model. Now let us turn to premise (2). To deny it, we should be able to say that the antecedent can be true, while the consequent is not. That is, whatever explains such a personal agentive contribution is not something other than the agent.

Recall what provides the values for each competing task, and what enables the agent to complete their internally maintained behavioural goals. It is not something which is completely alien to the agent, but rather, her psychological set-up: intentions, policies, and values. These are the structuring causes of her behaviour. To be sure, this account does not exclude the possibility that sometimes an agent can be alienated with respect to her standing mental states and the exercise of her cognitive control abilities, or that sometimes these abilities are not exercised in a way that is aligned to her standing states. However, at the very least, our proposal is not inconsistent with the possibility that in some cases, standing states and higher-order values constitutive of the agent are contributing to the setting of the utilities and biases that shape the allocation cognitive control resources. It is the agent that sets up the biases which guide action. The view of control presented here allows both for a mechanistic explanation of cognitive control and retains – and to some degree explains – the contribution of the agent in realising that control.

There are further reasons to find this philosophical interpretation appealing. Common intuitions about responsibility largely correspond well with the role of standing intentions and higher-order mental states in the cognitive control model. Recall Frankfurt's case of the unwilling addict (see Section 2). Presumably, the reason we are inclined to assign diminished responsibility to the unwilling addict is because her addiction is less an expression of her higher-order desires and self-governing policies. For the unwilling addict, biasing of attentional processing is not set in accordance with higher-order desires. The values that bias the unwilling addict's prioritisation of processing pathways involved in drug seeking and drug taking behaviour are computed independently of her higher-order goals. Therefore, she seems less responsible for her drug-related behaviour than the willing addict.

Cases like these indicate why we are often inclined to identify the agent and her contribution with higher-order mental organisation and the way this mental organisation influences control of thought and action. They also illustrate why the account presented here is in line with many compatibilist views on the connection between the agent and her mental states. Often, these views take the agent's higher-order mental states as constitutive of her psychological set-up. Responsibility intuitions seem to focus on this level, too. Providing an explanation of how that is psychologically realisable is exactly what the cognitive control model aims to do. We motivated the argument from mechanistic explanation with a possible explanation of the WCST. By explaining the participant's behaviour on ever lower levels, we seem less inclined to identify the mechanism with the agent. Now with an account of the structuring causal role of intentions and other standing

mental states, we have a way to avoid that problem. Important features of the computations at the lower levels refer back to the agent and her standing mental states, exactly the level at which intuitions about moral responsibility start to be effective.

Section 5: Two Objections

Before closing, we discuss two potential objections to our proposal. First, there is a worry about the degree to which our account overcomes the challenge to agentive contribution. While the cognitive control story seems to be consistent with agents guiding action through their standing goals or intentions, perhaps it only seems this way because the explanation is immature. When psychology has made more progress and has sufficiently matured, one will be able to provide a complete account of control. At this point the agent will again disappear from the psychological explanation. Once the details of the explanation are laid out fully, there will once again be no space left for the agent's contribution. This worry invites back the argument from mechanistic explanation, by questioning whether any proper scientific explanation of behaviour can be based on personal-level states such as agency.

A reply would start by acknowledging that the worry might in fact materialise. It may be that the type of psychological explanation we have described is premature. It is definitely lacking in detail. But at the same time, what determines when an explanation is mature enough? In order to respond to this question, we need an independent way of specifying the maturity of a scientific psychological account. What would such an independent criterion be? The answer is that we do not know. Absent that, however, we still need to choose between current competing accounts. In the case of the account presented in this paper, it is well confirmed by recent experimental and neuroscientific data, and it is computationally attractive.

The exact shape of the complete or mature psychological account is an open empirical issue. In the meantime, and until that empirical issue is resolved, we need a way to choose between the different accounts we currently have. If our best psychological evidence allows that the agent makes a real contribution in bringing something about, we have some reason to think that this represents psychological reality. At present, our best psychology of cognitive control provides us with an explanatory framework that is consistent with the claim that agentive contribution is possible. The form of explanation prevalent in the scientific literature is not inconsistent with a substantial role for persons as agents.

The second potential objection relates to the problem of the disappearing agent. The problem is associated with a general worry for event-causal accounts of action, which seems to indicate that any causal explanation of action ultimately leaves no space for the agent. Such accounts explain action in terms of causal relations between events, and there is no causal influence that can be attributed to the agent. The agent is reduced to the stage on which the chains of mental events run their course (Velleman, 1992; Pereboom, 2014).

The objection to our proposal is that since the challenge to agency posed by the argument from mechanistic explanation is like the problem of the disappearing agent for causal accounts of action, our proposal is susceptible to worries about disappearing agent like those used against causal accounts of action. One might complain that our proposal fails in dispelling the general problem of disappearing agent. Ultimately, to the extent that this objection is successful, the challenge posed by the argument from mechanistic explanation would collapse into the more familiar problem of the disappearing agent.

However, while both the argument from mechanistic explanation and the problem of the disappearing agent refer to a disappearance of agency, the two problems are significantly different. Whereas the challenge addressed in this paper concerns the disappearance of the causal role of personal level states in the mechanistic explanation of control in psychology, the problem of the disappearing agent is a general challenge to causal theories of action. To the extent that our proposal is a causal theory, we could be susceptible to any general objection to causal theories of action. But we could also avail ourselves of general replies. This would be a discussion for a different context. Our challenge and the general problem are different and separate issues. We are therefore not forced to provide a solution to the general problem of the disappearing agent.

Section 6: Conclusion

We began by establishing the relationship between agentive contribution and responsibility-level control. At least one way to settle whether an agent has the sort of control that is necessary for moral responsibility is to track the degree of her contribution. In fact, as cases like the willing and unwilling addict demonstrate, a diminished agentive contribution may be regarded as an excusing condition. Our primary question has been whether humans realise this kind of agentive contribution and thereby responsibility-level control. According to the argument from mechanistic explanation,

the form of explanations used in cognitive psychology and cognitive neuroscience leaves no space for agency, as personal-level capacities are explained by sub-personal mechanisms. Since most prominent theories of moral responsibility require that the agent plays some role in action in order to be morally responsible, the challenge motivates scepticism about the possibility of a psychological explanation of responsibility-level control. By integrating insights from theories of cognitive control and incorporating them into a philosophical account of the role of standing intentions as internally maintained goals and structuring causes of action, we have provided a way to think of the agent's contribution in functional-mechanistic terms. Standing intentions, long-term goals, and policies play a role in setting values and determining utilities central to the allocation of cognitive control. We have provided an account that is consistent with both the form of control necessary for moral responsibility and our best computational and mechanistic explanation of cognitive control.

References

Bermúdez, J. L. (2004). *Philosophy of psychology: A contemporary introduction.* Routledge.

Bermúdez, J. L. (2006). *Arguing for eliminativism*. In B. Keeley (Ed.), Paul Churchland: *Contemporary philosophy in focus* (pp. 32–66). Cambridge: Cambridge University Press.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology*, *66*, 83-113.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375-405.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (2004). Three theories of self-governance. *Philosophical Topics*, *32*(1/2), 21-46.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, *16*(2), 106-113.

Braver, T. S. (Ed.). (2015). *Motivation and cognitive control*. Routledge.

Buehler, D. (2019). Flexible occurrent control. *Philosophical Studies*, *176*(8), 2119-2137.

Buehler, D. (2021). Agential capacities: a capacity to guide. *Philosophical Studies*, 1-27.

Bugg, J. M., Jacoby, L. L., & Toth, J. P. (2008). Multiple levels of control in the Stroop task. *Memory & cognition*, *36*(8), 1484-1494.

Clarke, R. (2014). *Omissions: Agency, metaphysics, and responsibility*. Oxford University Press.

Cohen, J. D. (2017). Cognitive control: core constructs and current considerations. *The Wiley handbook of cognitive control*, 1-28.

Cummins, R. C. (1983). *The nature of psychological explanation*. MIT Press.

Davidson, D. (2001). *Essays on Actions and Events*. Oxford University Press.

Dennett, D. C. (1975). Why the law of effect will not go away. *Journal for the Theory of Social Behaviour*, *5*(2), 169-188.

Dretske, F. (1989). Reasons and causes. *Philosophical Perspectives*, *3*, 1-15.

Enç, B. (2003). *How we act: Causes, reasons, and intentions.* Oxford University Press.

Ford, A. (2018). The Province of Human Agency. *Noûs*, 52(3), 697-720.

Frankfurt, H. G.(1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5.

Frankfurt, H. G. (1988). *The importance of what we care about: Philosophical essays*. Cambridge University Press.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of experimental psychology*, *38*(4), 404-411.

Grünbaum, T. (2008). Trying and the Arguments from Total Failure. *Philosophia*, *36*(1), 67-86.

Grünbaum, T. (2009). Anscombe and Practical Knowledge of What is Happening. *Grazer Philosophische Studien*, 78, 41-67.

Grünbaum, T. (2012). Commonsense psychology, dual visual streams, and the individuation of action. *Philosophical Psychology*, *25*(1), 25-47.

Grünbaum, T. (2013). Seeing what I am doing. *Philosophy and Phenomenological Research*, *86*(2), 295-318.

Grünbaum, T., & Kyllingsbæk, S. (2020). Is Remembering to do a Special Kind of Memory?. *Review of Philosophy and Psychology*, *11*, 385-404.

Hornsby, J. (1980). *Actions*. Routledge.

Kaplan, D. (2017). Integrating Mind and Brain Science. A Field Guide. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science*, pp. 1-28. Oxford University Press.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation.* MIT press.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences*, *36*(6), 661-679.

Lavin, D. (2013). Must there be basic action?. *Noûs*, 47(2), 273-301.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

Mayr, E. (2011). *Understanding human agency*. OUP Oxford.

Mele, A. R. (1992). Akrasia, self-control, and second-order desires. *Noûs*, *26*(3), 281-302.

Mele, A. R., & Moser, P. K. (1994). Intentional action. *Noûs*, *28*(1), 39-68.

Mele, A. R. (2010). Moral responsibility for actions: Epistemic and freedom conditions. *Philosophical Explorations*, *13*(2), 101-111.

Mele, A. R. (2020). Direct Versus Indirect: Control, Moral Responsibility, and Free Action. *Philosophy and Phenomenological Research*. (forthcoming)

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological review*, *104*(1), 3-65.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733-7741.

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies*, 177(3), 825-843.

Nelkin, D. K., & Rickless, S. C. (2017). Moral responsibility for unwitting omissions: A new tracing view. *The ethics and law of omissions*, 106-129. Oxford University Press.

O'Shaughnessy, B. (2000). *Consciousness and the World.* Oxford University Press.

Pacherie, E. (2006). Towards a dynamic theory of intentions. *Does consciousness cause behavior*, 145-167.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179-217.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Pereboom, D. (2014). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies*, *169*(1), 59-69.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283-311.

Posner, M. I., & Snyder, C. R. R. (2004). *Attention and Cognitive Control.* In D. A. Balota & E. J. Marsh (Eds.), *Key readings in cognition. Cognitive psychology: Key readings* (p. 205–223). Psychology Press.

Schlosser, M. E. (2013). Conscious will, reason-responsiveness, and moral responsibility. *The Journal of ethics*, *17*(3), 205-232.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Shepherd, J. (2014). The contours of control. *Philosophical Studies*, 170(3), 395-411.

Shepherd, J. (2015). Conscious control over action. *Mind & language*, 30(3), 320-344.

Shepherd, J. (2016). Conscious action/zombie action. *Noûs*, 50(2), 419-444.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127-190.

Spencer, C. (2007). Unconscious vision and the platitudes of folk psychology. *Philosophical Psychology*, *20*, 309–327.

Sripada, C. (2017). Frankfurt's unwilling and willing addicts. *Mind*, *126*(503), 781-815. Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Steward, H. (2012). The metaphysical presuppositions of moral responsibility. *The Journal of ethics*, *16*(2), 241-271.

Stich, S. (1978). Autonomous psychology and the belief-desire thesis. *The Monist, 61*, 573–591.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Uithol, S., Burnston, D., & Haselager, W. F. G. (2014). Why we may not find intentions in the brain. *Neuropsychologia, 56*, 129–139.

Velleman, J. D. (1992). What happens when someone acts?. *Mind*, *101*(403), 461-481.

Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*(8), 205-220.

Watson, G. (1987). Free action and free will. *Mind*, *96*(382), 145-172.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227-248.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395-415.

Wu, W. (2016). Experts and deviants: The story of agentive control. *Philosophy and Phenomenological Research*, *93*(1), 101-126.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, *107*(3), 410-426.

Article 2:
The Worry about Mental Luck

Abstract

I argue that a problem of luck originally aimed at libertarians generalises beyond any specific theory of free will, and independently of whether determinism or indeterminism is true. I call this kind of luck mental luck. Mental luck suggests that, from the perspective of the agent, some decisions resemble the outcome of a lottery, caused by mental factors unknown to her. Mentally lucky decisions are not rationally controlled by attitudes within the agent's perspective, and they may be indistinguishable from non-lucky decisions. Therefore, mental luck poses a challenge to most prominent accounts of free action and moral responsibility.

Section 1: Introduction

There is a widespread view in both philosophy and common sense that moral practice should be immune to luck (see for instance Kant, 1784 [1998]; Nagel, 1979; Williams, 1981). Intuitively, it seems unfair to hold people responsible for what is not up to them or lies beyond their awareness. For instance, one could plausibly suggest that an action is free, and an agent is morally responsible for it only if its happening is in some meaningful way rationally controlled by attitudes which are reflectively accessible to the agent. Decisions that occur as a matter of luck typically do not fulfil even this broad and underspecified condition.

Luck is particularly problematic if one assumes that the universe is indeterministic because in that case, any event may be the result of chance. Any indeterministic happening is potentially the outcome of some chance process, determined by factors outside the agent. Therefore, an agent's decisions are lucky since they are explained by factors outside the agent herself. Holding people responsible for what they do seems untenable if what they do is a matter of luck.

In this paper, I argue that this problem of luck is not restricted to indeterminism. Regardless of whether we live in an indeterministic world some of our decisions are prone to a certain kind of luck, which I call mental luck. Mental luck affects agents regardless of the metaphysical nature of

the world. It has its origins in the limits of our mental life and how we make decisions under conditions of subjective uncertainty. I argue that the outcome of decisions that are prone to mental luck is not determined by the agent's accessible practical attitudes. And since most prominent theories associate the agent's moral responsibility with her practical attitudes and with the outcome of her decisions, mentally lucky decisions do not seem to be free and morally responsible. This conclusion points to a dilemma about the centrality of conscious awareness for the purposes of moral assessment. While it is possible to avoid the undesirable consequences of mental luck, doing so comes with its own set of problems.

The paper begins by presenting a version of the luck problem sometimes called *present luck* (Mele, 2006). The worry about present luck is typically introduced as a problem for libertarian theories of free will. Then, I discuss a proposal on how to generalise it beyond libertarianism. In section 3, I argue for a generalisable epistemic version of the problem of luck, which affects all agents regardless of whether the universe is deterministic or indeterministic. An agent may make different decisions, while her accessible mental states remain identical. In section 4, I consider how mental luck affects theories of moral responsibility. I finish the paper with a potential line of objection and a possible reply.

Section 2: The Worry about Present Luck

Luck usually seeps into the free will debate in the form of *moral luck*. Suppose an agent ends up in circumstance A, but that she could have just as well ended up in circumstance B. If the difference between ending up in A or B is a matter of luck and if the outcome has significance for purposes of moral assessment, then her ending up in situation A is a matter of moral luck. The term itself was introduced by Bernard Williams, who initially meant it as a paradox about common moral practice (Williams, 1981). Put simply, it seems counterintuitive to hold people morally responsible based on something that is the result of luck-related factors.

Thomas Nagel (1979) introduced a taxonomy of different kinds of luck that seems to influence the moral worth of agents in ways beyond their control. Nagel's taxonomy demonstrates how luck can affect the outcome of an act, the circumstances under which it occurs, or determine the agent's constitution or character. It also highlights the close relationship between moral luck and control. Understanding how luck affects free and morally responsible agents means understanding how luck

affects their level of control. Therefore, a discussion of the relationship between luck and moral responsibility is, essentially, a discussion about control.

At this point, it is important to distinguish between two types of control: *psychological-type control* and *responsibility-type control*. In psychology, control is often associated with a finite agential capacity to inhibit or resist interference while performing a task (Shiffrin & Schneider, 1977; Posner & Snyder, 2004; Cohen, 2017). Control refers to a limited capacity that agents possess, and they can apply it in different ways. However, in discussions of free will and moral responsibility, control is often understood as a necessary condition for moral responsibility (Zimmerman, 1988; Watson, 1996; Fischer & Ravizza, 2000; McKenna, 2008; Levy & McKenna, 2009; Björnsson & Persson, 2012; Levy, 2014a; Björnsson, 2017). By responsibility-type control I mean the type of control that a person must have to be blameworthy or praiseworthy. I leave it open how this control is related to the control as understood in the literature on cognitive control or action control. In this paper, I am interested in how luck affects agents in ways that undermine their responsibility-type control.

One particular luck-related worry which has a bearing on one's control of action was introduced by Mele in an attempt to pose a challenge for incompatibilist accounts of free will. Incompatibilism is the view that a deterministic universe is incompatible with the idea of free will. Determinism is in conflict the idea of free action, roughly because the truth of determinism suggests that the future is not up to us (van Inwagen, 1983). In a deterministic universe, incompatibilists argue, no one ever acts freely. Incompatibilism forces those who accept its tenets to choose between a deterministic universe without free will, or insist that we have free will, and thus, the universe is indeterministic. Incompatibilists who accept the first horn of the dilemma are commonly called hard determinists, while those incompatibilists who believe in free will are called libertarians.

Mele proposed the worry about present luck in an effort to demonstrate a problem with libertarian accounts of free will. In contrast to Nagel's taxonomy, this way of describing moral luck applies not to results or causes, but to the decision-making process itself. Present luck occurs in a scenario which captures an agent's decisions in two possible worlds that share the same past and the same laws of nature. The worry was originally formulated as such:

> *The Worry about Present Luck (PL): If an agent freely decided at t to A, then in a possible world with the same past until t and the same laws of nature, he did not decide at t to A* (Mele, 2006: 58).

Now let us consider how present luck presents a challenge for libertarian accounts. At first glance, cases like the one above seem paradoxical. How can an agent freely decide at t to A, and yet in a possible world where nothing about the agent has changed, not decide at t to A? Yet, according to libertarianism, such cases are possible. Because for the libertarian the universe is indeterministic, she must accept that, at t, it is open to the agent whether he will decide to A or not decide to A. Furthermore, according to the libertarian, deciding at t to A is an instance of free and morally responsible agency. The task is to explain the difference between deciding at t to A and not deciding at t to A. Since in the formulation of present luck the past and the laws of nature are fixed, there is nothing in the past of the agent or in the way the two worlds are composed that can explain that difference. Therefore, the difference between deciding at t to A and not deciding at t to A is a matter of luck.

To see this more clearly, consider the following conceptualisation of the two possible worlds in cases of present luck. The worlds are correspondingly described by the two cases below:

> *Basic Case (BC): An agent decided at t to A.*
> *Alternative Case (AC): An agent did not decide at t to A.*

According to the worry about present luck, an agent can end up in either BC or AC. The past and laws of nature are fixed, and there is nothing about the agent that explains the difference between ending up in BC or AC. The libertarian finds herself in an untenable position: the occurrence of a seemingly free decision occurs is a matter of luck.

Libertarians have developed several replies to the worry about present luck. Perhaps the most prominent of those replies is Robert Kane's account of self-forming acts, according to which both deciding and not deciding at to A are voluntary, free, and responsible actions (Kane, 1998; 1999; 2006). Other replies adopt a more agent-causal explanation, such as the ones proposed by Randolph Clarke (2004; 2005; 2011) and Timothy O'Connor (2007; 2011). The details of those replies are not essential to the present argument, apart from the fact that all of them attempt to stand up to present luck on metaphysical grounds. That is unsurprising, since present luck is itself a metaphysical

challenge. However, I believe there is a non-metaphysical way to understand cases of present luck. By non-metaphysical, I mean a way that does not depend on whether the universe is deterministic or indeterministic.

Before I review existing ways to generalise present luck and move on to my own proposal, it is important to briefly consider the relationship between present luck and indeterminism. Indeterminism is the doctrine that not all events are wholly determined by antecedent causes. Libertarian accounts of free will assume indeterminism because the sort of free will that libertarians endorse is only possible if some events are not wholly determined. Only in an indeterministic universe is it open at t whether the agent will decide to A or not. In other words, indeterminism has present luck built into it. And since present luck can apply to morally significant decisions, such decisions would be subject to moral luck.

With present luck established as a serious challenge to those accounts of free will that assume indeterminism; the question remains of whether the challenge can be generalised beyond those accounts. If it cannot, the worry is cogent, but its scope is restricted to libertarianism. But if it can be separated from the truth of indeterminism and retain its force, present luck would affect agents regardless of the metaphysical structure of the world. Such a generalisation would suggest that present luck is not merely an issue for incompatibilism, but also for compatibilism, namely the thesis that determinism is compatible with free will.

In recent years, there have been a few voices in that direction (see for instance Vargas, 2012; Perez de Calleja, 2014). In this paper, I consider the second of those arguments. Perez de Calleja claims that present luck presents a problem for compatibilism, not because the decisions in question are indeterminate (as that wouldn't be a problem), but because the agent in question is motivationally split. Whatever makes the agent decide as she does cannot be explained by her mental states. This not a consequence of indeterminism, but of the agent's diverging motivations. Such lucky and motivationally split decisions are equally problematic for incompatibilists and compatibilists alike.

Perez de Calleja turns the focus on how luck-prone decisions appear for the agent. Furthermore, she formulates the problem in terms that do not presuppose indeterminism. Both these elements will also form part of my own proposal to generalise present luck beyond libertarian accounts of free will. However, her account is conditional on a very restrictive view of deterministic

luck, namely that luck-prone decisions are motivationally split, in the sense that the agent is in some meaningful way divided between different decisions. So, while deterministic luck extends beyond indeterminism, it seems limited to decisions that are motivationally split. Thus, it does not seem to affect decisions that are the product of more stable mental states, such as policies, or decisions that are inextricably linked to the agent's plans and future-directed intentions.

Perez de Calleja is right that agents may be prone to luck that is not a result of indeterminism. However, such luck does not need to be a result of determinism, either. The account is overly restrictive on this point. It is not that agents are subject to luck determinism is true because their motivations are split. Rather, they are subject to luck because of some epistemic constraints on their psychology, which affect the way they deliberate and make decisions. As I want to argue, present luck indeed generalises, independently of the metaphysical nature of the world.

Such a metaphysically independent version of present luck would extend the worry to influence potentially any decision. With that in mind, in Section 3 I move away from the existing discussion on present luck. I also move away from the discussion of luck within the confines of the free will debate, and towards a different and more generalisable version of the problem. I propose a version of the worry which associates moral luck with something more general, namely the epistemically limited nature of the agent's perspective, and the significance such limitations have for one's decisions.

Section 3: The Worry about Mental Luck

3.1 Introducing Mental Luck

In the following section, I present a different way to generalise a version of present luck beyond libertarianism, and independently of one's metaphysics. This section attempts to make sense of the problem of luck, this time from the perspective of the agent. What would an agent's experience be like, if she were to find herself in a situation like the one that motivated the worry about present luck?

Let's begin by revisiting the two cases which describe the different possible outcomes in instances of present luck:

Basic Case (BC): An agent decided at t to A.
Alternative Case (AC): An agent did not decide at t to A.

In this section I will argue for the following claim: In cases like BC and AC, there is always some probability that the difference between deciding at t to A and not deciding at t to A is matter of luck from the perspective of the agent. In other words, from the perspective of the agent, there is a chance that whether she decides at t to A resembles the outcome of a lottery. This conclusion allows luck to seep into potentially any decision. Ostensibly, the probability of this happening is low, but importantly, it does not depend on whether the universe is deterministic or indeterministic.

To express that probability in more familiar terms I introduce a novel version of present luck. For reasons that will become clear I call this novel version "mental luck":

> *Mental Luck: If an agent freely decided at t to A, then in a possible world* where she has the same perspective *until t, she did not decide at t to A.*

Before introducing mental luck in more detail, a closer look on how mental luck differs from present luck. In Mele's original formulation, what is kept constant is the past and the laws of nature. By contrast, in this epistemic version, what is kept constant is the epistemic outlook of the agent (her perspective), while allowing that the past can change. The laws of nature are obviously unaltered.

As with present luck, the force of the worry depends on the difference between BC and AC being a matter of luck. Or, more specifically: From the perspective of the agent, there is a chance that whether she ends up in BC or AC is a matter of luck. I argue that from the agent's perspective this possibility cannot be excluded. Furthermore, I argue that the worry about mental luck is in fact generalisable beyond libertarianism. Mental luck captures something that is true of agents regardless of one's metaphysical requirements for free will.

In order to support these claims, I begin by construing the notion of perspective and its relation to accessibility of mental states. Then, I elaborate on two elements of mental luck that differentiate it from present luck: *metaphysical independence* and *perspective identity*. With these pieces in place, mental luck emerges as a worry both akin and different to present luck. From there, I move on to a discussion about *knowledge* and *justification*, which demonstrates how the truth of certain

beliefs may be a lottery to the agent. That same discussion also provides additional reasons to think that the perspective of the agent can remain constant between physically different possible worlds, as in BC and AC. Finally, I conclude that what is true of certain beliefs, namely that their truth may seem like a lottery to the agent, can also apply to decisions. This conclusion indicates that mental luck challenges free and morally responsible decisions independently of the metaphysical structure of the world.

3.2 Perspective and Accessibility

Mental luck is reminiscent of present luck, but it is also quite different from it. One way it differs is regarding the centrality of the notion of perspective. Mental luck is motivated largely by the suggestion that the perspective of the agent can remain constant, whether she ends up deciding at t to A or not. Before I can discuss why that is in fact possible, I need to say what I mean by perspective.

For the purposes of construing the notion of perspective as it features in the worry about mental luck, I make two assumptions. First, that perception should typically provide reasons for agents to make rational decisions in relation to the external world. That is, the agent's perspective consists in mental content that is informed by her perception, and that these states are playing a rational role in her interactions with the world. Second, that the agent does not have guaranteed access to the entirety of the objective world. This incompleteness of access refers to both states of the external world and to some of her own mental states. I discuss this second assumption in more detail further down.

Consequently, the agent's perspective does not include all of the psychological facts about her. From this we can begin to pick out which mental states are in fact included in the agent's perspective. We start with the obvious: mental content that is consciously accessible to the agent. Such mental states are definitely part of one's perspective. Perhaps even states of which she is not explicitly conscious can be included, as long as they are somehow accessible to her in the right way. At any rate, a mental state is within an agent's perspective only if that state is conscious or open to the right kind of reflection. And since it is reasonable to assume that there are yet other states on which the agent cannot readily reflect, or states of which she is not consciously aware, the perspective does not include all of the psychological facts about the agent.

There are examples of the contrast between mental states within or outside the agent's perspective. For instance, consider the difference between conscious and non-conscious perception. Agents are constantly presented with all sorts of perceptual content from the external world. Some of this content is conscious: the agent is aware of it, she relates it to the rest of her mental states, and it plays a role in her rational thinking, deliberation, and action. She has reflective access to the content. In some rough but reasonable way, such content makes sense to the agent. But some perceptual content is likely not conscious. Notwithstanding, such non-conscious perception might influence the agent's thinking, deliberation, and action in all sorts of ways. The influence of such non-conscious content is not reflectively accessible to the agent.

The extent to which an agent's mentality lies within or outside her perspective, and whether the line between the two is a matter of degree, are fascinating questions. I do not presume to know the answers to them. For my purposes, what is important is that mental states outside the agent's perspective influence her actions in various ways, but they are not accessible to the agent. Because of that, such states do not play a straightforward role in rational thinking and deliberation. So, it is possible for something that lies outside the agent's perspective to influence her conduct in ways that do not line up with those mental states that form her perspective.

Exactly which mental states can play this role and how often it occurs is an issue I take up later in this section. For now, it suffices to say two things: First, the agent's perspective includes mental states that are accessible to her by means of reflective awareness. Second, the agent has mental states that do not form part of her perspective. Next, I move on to two defining elements of the worry about mental luck that set it apart from the discussion on present luck. These are metaphysical independence and perspective identity.

3.3 Metaphysical Independence and Perspective Identity

The first element of mental luck that distinguishes it from present luck concerns metaphysical independence. The worry about mental luck is meant to be independent of one's metaphysics and as such, generalisable beyond just libertarianism. To determine why that is the case, we need to consider whether the proposed formulation of mental luck does in fact steer clear of such discussions, or if it is instead caught up in the same metaphysical discussion around the worry about present luck.

Regarding present luck, the original formulation of the worry entails indeterminism. Present luck arises because in libertarian accounts, it is open whether the agent ends up in a world where BC or AC obtains, up until t. The two worlds are identical, up until the moment of her decision, and yet she is free to decide to A or not decide to A. So, present luck poses a challenge for indeterministic accounts of free will, but less obviously so for views that do not require indeterminism.

Now to mental luck. The worry suggests that the past (physical and psychological) leading up to BC may be different from the past leading up to AC. From the perspective of the agent, however, that difference is not accessible. Up until t, the two worlds are indistinguishable from her perspective. Note that mental luck applies to both a deterministic and an indeterministic world. So, to the extent that mental luck presents a challenge, it does so without committing to any metaphysical account. As a result, if it is a challenge at all, it is a challenge to all theories of free will.

Across each of the two possible outcomes (BC and AC), the agent's perspective remains unaffected. But is this possible? Can the agent's reflectively aware, accessible mental states be identical, whether she decides to A or not? Consider how this differs from the worry about present luck. Assuming supervenience, if the physical world remains constant, so does the agent's psychology. But in cases of mental luck, it remains open that the physical or psychological world may vary. Mental luck does not commit us to inter-world variance, but it allows for it.

Now, if the perspective remains constant, would this imply that the physical world remains constant too? After all, identity of perspectives implies that the agent's reflectively aware and accessible mental states will also be identical. Wouldn't that imply that the physical world is also constant? If that were the case, identical perspectives would entail identical physical worlds, and mental luck would end up being indistinguishable from present luck. There are several ways around this worry. One may accept some notion of supervenience of the mental upon the physical. Such supervenience would suggest that there cannot be a difference in the mentality of the agent without a difference in the physical world. Correspondingly, physical world identity would imply mental world identity. However, the reverse does not need to be true. Mental world identity does not necessarily entail physical identity. This would allow for multiple realisability, namely the thesis that a mental kind (such as a belief or a desire) can be realised by multiple physical kinds.

Coming back to the worry about mental luck, identical perspectives in BC and AC do imply mental identity of perspectives. But if multiple realisability is true, mental identity does not imply physical identity. In fact, the fact that there is some mental identity of perspectives tells us very little about the physical world of BC and AC. It is perfectly possible for the physical world to differ between BC and AC, while the mental world of the agent remains unaffected, just as the worry about mental luck suggests.

Multiple realisability provides a basis that allows variance between physical worlds, while at the same time maintaining that the agent's perspective remains constant. But what if the thesis of multiple realisability is false? What if, instead of multiple realisability, something like identity theory is true? That would mean that identical mental worlds entail identical physical worlds. Here we have a second way to tackle the worry. Mental states that form the agent's perspective are only a subset of her total mentality. If identity theory is true physical variance entails mental variance. Any change in the physical world corresponds to changes in the agent's total mentality. But since her total mentality includes more mental states than those to which she has access, changes in total mentality do not entail changes in the agent's reflectively aware, accessible mental states which form her perspective. All this can be summed up slogan form: identical perspective does not entail total mental identity. That leaves open the possibility that, even when the agent's mentality changes along with the physical world, her perspective remains identical.

3.4 Knowledge and Lottery

With the notion of perspective laid out, it remains to be seen whether ending up in BC or AC is a matter of luck from the perspective of the agent. For reasons that will be become clear, in this next part I depart from the formulation "a matter of luck" and settle into using "lottery" instead. While luck and lottery can mean very different things in various contexts, I take them to be equivalent for my purposes.

Expressing mental luck in terms of a lottery between two outcomes (deciding to A and not deciding to A) has relevant similarities with a well-developed and relevant body of work from epistemology. This familiar debate about the conditions under which a belief constitutes knowledge provides a way to better understand decisions under the influence of mental luck.

Epistemic luck occurs when an agent has a justified true belief, but that belief is too lucky to count as knowledge (for an introduction to epistemic luck see Latus, 2000; Pritchard, 2005; Coffman, 2015). Typically, the departure point is found in Gettier cases (Gettier, 1963; for an overview see Hetherington, 2011), which demonstrate that it is possible for an agent to hold a belief that is both justified and true, but arguably, does not constitute knowledge. From the agent's perspective, whether her belief is true is a lottery. Importantly, the truth maker is external to the agent. Whether the belief is true is a lottery because there is nothing about the agent that can help her determine its truth. In order to see this more clearly, consider the following statements about epistemic luck:

> *(1) If there is nothing in the agent's perspective that distinguishes between her believing that P is true and not believing that P is true, then whether P is true is a lottery from the perspective of the agent.*
>
> And also:
> *(2) In cases of epistemic luck, there is nothing in the agent's perspective that distinguishes between believing that P is true and not believing that P is true.*

Therefore, from (1) and (2), in cases of epistemic luck, whether P is true is a lottery from the perspective of the agent.

Now consider a corresponding line of argument about mental luck. As epistemic luck suggests, the agent may lack access to the truth value of a belief which happens to be true. The truth of that belief is therefore a lottery from the perspective of the agent. Mental luck focuses on decisions. In cases of mental luck, the agent may lack access to whatever makes the difference between deciding at t to A or not deciding at t to A. This idea can be formulated in a similar statement to (1) as follows:

> *(3) If there is nothing in the agent's perspective (up until t) that distinguishes between her deciding at t to A (BC) and not deciding at t to A (AC), then whether she ends up in BC or AC is a lottery.*

Let us look at the mental luck version of (1) more closely. (3) is an expression of the idea that as long as the agent does not distinguish between deciding at t to A and not deciding at t to A, then which of these two outcomes she ends up in is a matter of luck.

Now, I have argued that in cases of mental luck the perspective of the agent up until t can remain identical. This idea can be expressed in the following:

> *(4) In cases of mental luck, there is nothing in the agent's perspective (up until t) that distinguishes between her deciding at t to A (BC) and not deciding at t to A (AC).*

It follows from (3) and (4) that in cases of mental luck whether an agent decides at t to A or not is a lottery from the perspective of the agent. Having said that, nothing about the way such a belief is formed is necessarily problematic or particularly unusual. Given all the facts about the agent and her knowledge of the world, it may well be justified or rational to hold it. But the truth of such a belief is a lottery. In other words, the fact that some belief is not only justified but also true, may be a lottery.

3.5 Mental Luck as a Lottery of Decisions

In the epistemic luck literature, lucky events can be described in either modal (Pritchard, 2014), or probabilistic terms (Rescher, 2019), or a combination thereof (see de Grefte, 2020, for a hybrid account). Each of these accounts captures some aspects of how luck seems from the perspective of the agent. But what does that tell us about mental luck, and the possibility that a decision is a lottery, from the agent's perspective? After all, the worry about mental luck is quite a departure from the epistemic luck debate, simply because the mental luck worry is not concerned with knowledge and belief. The truth of a belief may be a lottery because its truth maker is external to the agent. But the same cannot be said of cases like those that motivate the worry about mental luck. Here, there is no external decision-maker. Whatever makes the agent decide or not decide at t to A is internal to the agent.

An agent ends up in BC when she decides at t to A. Her decision is explained by some combination of beliefs and desires. Her perspective is a subset of her total mentality, and it includes accessible, potentially reflectively aware mental states. Within these mental states, there are the beliefs and desires that explain her decision at t to A. Her decision is comprehensible to her. In

another possible world, the agent ends up in AC. Similarly, the fact that she does not decide to A is explained by some combination of beliefs and desires. However, at least a part of that combination lies outside her perspective. Suppose P is a belief that partly explains the fact that the agent does not decide at t to A. P is part of the agent's mentality and influences her conduct in various ways. But the agent is in a meaningful way unaware of that since P is not an accessible mental state. The fact that she does not decide at t to A may be explained by P, but she has no way of knowing that. Her decision is puzzling to her, perhaps incomprehensible.

Up until t, those two possible worlds may be physically or psychologically distinct in all sorts of ways. But from the perspective of the agent such difference is not accessible since her perspective remains constant across the two possible worlds. There is a possibility that an agent can have the same combination of reflectively aware and accessible beliefs and desires, but still end up in either BC or AC. From her perspective, the two worlds are identical, but in the first she decides to A, while in the second she does not decide to A. In the world of AC, whatever combination of beliefs and desires keeps her from deciding to A is not part of her accessible mental states. It is external to her perspective. So, while mental luck is a problem internal to the agent, the agent's decision is caused by mental states that are external to the agent's perspective.

Ending up in AC is mental bad luck, much like in cases of epistemic bad luck. It is bad luck because it goes contrary to her accessible mental states. When the agent ends up in AC, her decision is determined by some inaccessible mental state outside her perspective. But correspondingly, since the perspective remains constant, ending up in BC is also a matter of luck. The outcome might be less puzzling to her, but she could have just well not decided to A. So, from the agent's perspective, whether she ends up in either BC or AC is a lottery.

3.6 Implicit Attitudes

What kind of mental state might remain largely inaccessible and yet drive the agent towards AC? Particularly good candidates are implicit attitudes. Typically, implicit attitudes are largely unconscious, introspectively unidentified associations between concepts on the one hand and evaluations on the other (Fazio et al., 1995; Greenwald et al., 2009). Apart from being unconscious or inaccessible to the agent, they also mediate judgements and actions towards social objects in favourable or unfavourable ways (Greenwald & Banaji, 1995; Banaji & Greenwald, 1995). So, for instance, one may harbour some implicit attitude against young children and, thus, make decisions

that are unfavourable to this particular group, without being aware of the fact. This feature of implicit bias is often called *introspective opacity* (Washington & Kelly, 2016). Introspective opacity suggests that implicit attitudes are particularly resistant to introspective reflection and therefore to mediation or regulation.

Another feature that is relevant for our purposes is the observed disparity between explicit and implicit attitudes. People's explicit convictions can sometimes contradict their implicit attitudes, the latter exemplified by their behaviour (Yovel & Friedman, 2013; Teachman et al., 2008). One may explicitly hold egalitarian beliefs and yet act in ways that are unequivocally non-egalitarian. While measuring this disparity is often difficult, the finding seems to capture a common intuition: among other things, our actions may be the expression of some unconscious, reflectively inaccessible belief. The literature on implicit biases is both extensive and much debated. For now, it suffices to say that there seem to be some attitudes which can mediate decisions in ways that contradict the explicit beliefs of the agent, and that these attitudes are -at least sometimes- beyond the agent's introspectively available awareness.

Bringing the conversation back to familiar terms, epistemic luck is a worry about a belief, the truth of which is a matter of luck. Correspondingly, mental luck is proposed as a worry about a decision which, from the perspective of the agent, is not determined. Or, more precisely, it may well be determined, as the worry is not concerned with the metaphysics of such cases. But it is not determined by anything within the agent's perspective, by her accessible beliefs and desires. Such a risk might be small, but still it exists. There is always a probability that whether she ends up in BC or AC is a lottery to her.

Of course, real agents usually have access to a lot of relevant information about the world, and they are typically capable of planning their decisions and actions effectively, according to the knowledge they possess. One may reasonably argue that rarely does an agent actually find herself in such a predicament. But the point of the discussion here is to demonstrate that, no matter how rich in knowledge the agent might be, from her perspective, there is always a chance she is missing some important piece of information. Something that is inaccessible to her, but makes the difference between ending up in BC or AC.

Section 4: Mental Luck and Moral Responsibility

So far, I have argued for the worry about mental luck, a version of present luck where the perspective of the agent remains constant, while the physical world may differ. Agents subject to mental luck may have access to identical mental states, and yet end up either deciding at t to A (BC) or not deciding at t (AC). From the perspective of the agent, whether she ends up in BC or AC may be a lottery, and the difference between them a matter of mental luck. Such decisions under the influence of mental luck are not rationally controlled by attitudes in the agent's perspective. Before closing, I consider the importance of mental luck for theories of free will and moral responsibility.

Are mentally lucky decisions free and morally responsible? One obvious reply is no. Typically, lucky decisions are beyond the agent's rational control. That would suggest that mental luck is a sort of excusing condition, if one belongs to arguably the majority of those who think control is a necessary condition for moral responsibility. Furthermore, since mental luck is independent of whether the universe is deterministic or indeterministic, and since it can potentially affect all agents, mental luck appears as a general worry against all theories of free will and moral responsibility.

However, things are more complicated. For once, consider exactly how decisions under mental luck are lucky: they are the outcome of a lottery from the perspective of the agent. One could say that it is only when we limit ourselves to the perspective of the agent that mental luck is an issue. But it is a separate question whether limiting our scope to the perspective of the agent is necessary for moral assessment. If the answer is yes, then we have reason to think that prominent theories may wrongly attribute blame or praise for mentally lucky decisions. If the answer is no, assessing the moral responsibility of an agent only on the basis of her perspective is inadequate.

Let us now take stock: if mental luck undermines moral responsibility, it does so because agents are held responsible not for the totality of their mental states, but for those states that form part of their perspective. Are there theories that support this claim? I think the answer is yes. At least for some theories of moral responsibility, the fact that a mental state is inaccessible and outside the agent's perspective is an undermining factor for ascribing blame or praise for actions that stem from that mental state (see for instance Levy 2014a; 2014b) on *access consciousness*, Alex Madva (2019) on *phenomenal consciousness*, or Harry Frankfurt (1971) on the *reflectively aware, practically reasoning agent*). For such theories, deciding to A is a morally responsible decision only if it is rationally controlled by attitudes within the agent's perspective.

In fact, the centrality of the agent's perspective in moral assessment brings out a dilemma for moral responsibility theories. On the one hand, we have theories which accept the perspective of the agent as central for moral assessment, like the ones mentioned above. Arguably, they represent the majority view. Regardless, these theories face the issue of mental luck. I elaborate on that in the remainder of this section. On the other hand, we have theories of moral responsibility which deny the centrality of the agent's perspective for the purposes of moral assessment (for instance George Sher's (2009) account of responsibility without awareness). Such theories avoid the mental luck issue, but they face a host of other problems. For instance, they struggle to explain the importance of ownership and control, the intentional component of action, or common intuitions about the significance of consciousness in moral assessment, to name just a few. For those who accept the central role of the agent's perspective in attributing moral responsibility, mental luck is a problem.

Consider the case of implicit attitudes. In the previous section I nominated them as one kind of mental state that might give rise to mental luck. Implicit attitudes lie, almost by definition, beyond the agent's perspective. Recently, there has been a growing literature on whether implicit attitudes undermine moral responsibility (Holroyd, 2015; Brownstein, 2016; Washington & Kelly, 2016; Levy, 2017; Madva, 2019). Some, like Neil Levy, go as far as claiming that empirical evidence on the effects of implicit attitudes provides reasons to be sceptical about the blameworthiness and praiseworthiness of agents in general. There is still disagreement on whether people are responsible for their implicit attitudes. But for our purposes, it is important that these attitudes pose a challenge to moral responsibility, precisely because they lie outside one's perspective.

Now, there may be theories that fare better than others when it comes to assigning blame for these largely unconscious and inaccessible attitudes. For instance, attributionists like Angela Smith (2008; 2015) and Matthew Talbert (2016) argue that agents may be morally responsible for such attitudes even when they are outside their direct control, because they reflect their deeply held views or evaluative judgements. However, still for such views, a version of the problem persists. Arguably, non-conscious attitudes lack judgement sensitivity of the right sort. It seems problematic to hold people responsible for their unconscious attitudes, especially if they are not sensitive to judgement. In any case, for the majority of philosophers -compatibilists and incompatibilists alike- control of some sort is necessary for moral responsibility. For them, holding people responsible for their implicit attitudes is challenging, exactly because these attitudes are outside the agent's perspective.

In sum, mental luck raises a challenge for prominent moral responsibility theories. Those theories that accept the centrality of the agent's perspective for the purposes of moral assessment find in mental luck a serious challenge. Theories that deny the perspective of the agent as central face several other difficult issues. Mental luck may not be a universal head-on problem, but it poses a challenge that is not easily avoidable.

A final note on the professed rare occurrence of cases of mental luck. As I mentioned in the previous section, one could reasonably expect cases of mental luck to be few and far between. So, if mental luck is so exceptional, why should it challenge our everyday conditions for moral assessment? In reply, the challenge laid out here does not rest on mental luck being a frequent occurrence. Rather, it rests on the fact that any decision could be the outcome of a lottery, unbeknownst to the agent. Mental luck is not a problem because it affects all agents all the time. It is a problem because it may affect any agent, at any time.

Section 5: Objections

While there may be various objections to mental luck, these can be reasonably incorporated into a single counterargument. At its core, this objection challenges a central premise of the argument for mental luck, namely the invariance of the agent's perspective between BC and AC. Defenders of the objection might argue that as long as the decision outcomes in BC and AC are different, surely the perspective of the agent leading up to them must be different, too. To accept that someone may end up deciding or not deciding at t to A while her perspective remains constant, is akin to insanity: ordinary agents simply do not act this way.

A way to flesh out the objection suggests that BC and AC cannot be the same from the perspective of the agent, because they cannot contain the same mental states. More specifically, if the agent's decision at t to A is the result of some combination of beliefs and desires, then the belief-desire that leads to BC is different from that which leads to AC. And since the belief-desire is different, so is the agent's perspective leading up to t. In short, different decisions mean different belief-desire combinations, and different belief-desire combinations mean different perspectives. Therefore, the invariance of the agent's perspective up until t is indefensible. This argument seeks to prevent the separation of decisions from the combination of beliefs and desires that led to them.

Decisions have causes, and these are typically found among the agent's beliefs and desires. Furthermore, some decision may be causally dependent upon some combination of beliefs and desires. However, even if the belief-desire that led to BC is different from the belief-desire that led to AC, there is no guarantee that those beliefs and desires that make up the difference between BC and AC are indeed part of the agent's perspective. So, while the objection may convey something true about the relationship between belief-desire and decision, it does not exclude the possibility that different decisions may be explained by different belief-desire combinations, which are in turn not fully part of the agent's perspective. As a result, the way to achieve the professed invariance of perspective is still open.

Now to a different way to formulate and motivate the objection. Suppose that an agent ends up not deciding at t to A, when from her perspective she could have just as well decided at t to A. Suppose further that within her perspective, deciding at t to A makes sense. A-ing seems perfectly rational to her, and it aligns with the rest of her mental states. Admittedly, there is something strange about this agent. To the extent that deciding at t to A makes sense to the agent, ending up in AC where she does not decide at t to A must strike her as puzzling. But does that mean that her perspective cannot remain identical, whether she decides to A or not? The answer to that question may depend on how one thinks about rationality, and the way that an agent makes up her mind about what she has most reason to do.

The defender of the objection might begin by pointing out that the sort of knowledge that we have about ourselves is special in at least two ways (Moran, 2002). First, agents have special access to their own mental states, such that they can know their own beliefs and attitudes in an immediate way. Second, judgements about oneself have special authority, such that they have a prima facie claim to truth. If self-knowledge means special access and authority over one's mental states, then not deciding to A seems absurd. Only an agent that is akratic, or perhaps suffers from a condition that causes her to procrastinate indefinitely, would be a likely candidate for mental luck. Since the mentally lucky agent is not in principle suffering from any condition, one way to understand the objection is to argue that AC is a case of akrasia. If A-ing makes sense to the agent, not deciding to A goes against her better judgement, thus making AC a case of akratic action.

In order to see how this line of argument gets traction, consider the following two possible interpretations of an agent who ends up in AC. The first interpretation assumes a more internalist approach to rationality (see for instance Moran, 2002; 2012). According to internalism, an akratic

agent judges best to A, then decides to B. In cases of mental luck, the internalist sees an agent who reaches a practical conclusion (A-ing), yet proceeds to not decide to A. This resembles a kind of akratic self-deception. However, mental luck cases are different in a significant way. In BC, the agent reasons and decides to A. In AC, the agent may reason similarly, but she does nothing. There is no decision, and no action. AC is more akin to an open-ended procrastination than to self-deception or akratic action.

Now to an alternative interpretation of AC. For this, suppose a more externalist view of rationality (for such a view see Arpaly, 2000). According to that view, agents are not necessarily irrational when they do not act according to their best judgement. It may be more rational -or at least less irrational- to act against what one judges to be the best course of action. Under an externalist approach, it is not irrational to end up in BC and AC while retaining the same perspective. In AC the agent may well act against her judgement when she does not decide to A, but she may still be rational. Not deciding to A is puzzling to her, but it may still be an instance of rational conduct.

Overall, on an externalist approach to rationality the agent's perspective can remain identical between BC and AC without necessarily ending up in cases of akratic or irrational action. But even if one assumes internalism about rationality, AC is not a case of akrasia. Rather, it is a case of open-ended procrastination. Such procrastination can go on forever and admittedly, that would be very puzzling. But it is not irrational. After all, the agent does not make any decision that contradicts her practical judgement. She simply decides nothing. As long as her conduct is lucky and as long as it has moral significance, the challenge of mental luck for theories of moral responsibility remains intact.

Before closing, a final remark. Objecting that AC is akin to akrasia captures something true about mental luck. In AC, the agent seems to be acting irrationally. That is especially true the more internalist about rationality and practical reasoning one is, and when one takes into account only those mental states that form part of her perspective. However, the question is less whether the mentally lucky agent is irrational and more whether cases of mental luck occur. I think that they do. It is true that if an agent reaches a full-blown practical judgement that she should A, and then proceeds not to decide to A, there is no space for mental luck. Such an agent is either akratic, delusional, under the influence of self-deception, or otherwise utterly irrational. However, mental luck is neither motivated nor challenged by such cases. Rather, in arguing for mental luck, I am

interested in cases of practical deliberation that are less clear-cut and more representative of our everyday experience in the world. Perhaps not too much depends upon A-ing or not. Perhaps the difference between the two is insignificant, or the possible repercussions are unknown to the agent. Perhaps there is a lot to consider for the agent, so she makes a decision under conditions of relative uncertainty and without having considered every accessible piece of information. Such are the practical judgments that are prone to mental luck.

References

Arpaly, N. (2000). On acting rationally against one's best judgment. *Ethics*, *110*(3), 488-513.

Banaji, M. R., & Greenwald, A. G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology*, *68*(2), 181–198.

Björnsson, G., & Persson, K. (2012). The explanatory component of moral responsibility. *Noûs*, *46*(2), 326-354.

Björnsson, G. (2017). Condition on Moral Responsibility. *Responsibility: The epistemic condition*.

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, *7*(4), 765-786.

Clarke, R. (2004). Reflections on an Argument from Luck. *Philosophical Topics*, *32*(1/2), 47-64.

Clarke, R. (2005). Agent causation and the problem of luck. *Pacific Philosophical Quarterly*, *86*(3), 408-421.

Clarke, Randolph (2011). Alternatives for Libertarians. In Robert Kane (ed.), The Oxford Handbook of Free Will, 2nd edition. pp. 329-48.

Coffman, E. J. (2015). *Luck: Its nature and significance for human knowledge and agency*. Springer.

Cohen, J. D. (2017). Cognitive control: core constructs and current considerations. *The Wiley handbook of cognitive control*, 1-28.

de Grefte, J. (2020). Towards a Hybrid Account of Luck. *Pacific Philosophical Quarterly*, *101*(2), 240-255.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68, 5-20.

Gettier Edmund, L. (1963). Is Justified True Belief Knowledge?. *Analysis*, *23*, 121-123.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4-27.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

Hetherington, S. (2011). The gettier problem. In *The Routledge companion to epistemology* (pp. 145-156). Routledge.

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and cognition*, *33*, 511-523.

Kane, R. (1998). *The significance of free will*. Oxford University Press on Demand.

Kane, R. (1999). Responsibility, luck, and chance: Reflections on free will and indeterminism. *The journal of philosophy*, *96*(5), 217-240.

Kane, R. (2006). Libertarian accounts of free will.

Kant, I. (1784) [1998], *Groundwork of the Metaphysics of Morals*, M. Gregor (ed. and transl.), Cambridge: Cambridge University Press.

Latus, A. (2000). Moral and epistemic luck. *Journal of Philosophical Research*, *25*, 149-172.

Levy, N., & McKenna, M. (2009). Recent work on free will and moral responsibility. *Philosophy Compass*, *4*(1), 96-133.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

Levy, N. (2014). The value of consciousness. *Journal of Consciousness Studies*, *21*(1-2), 127-138.

Levy, N. (2017). Am I a racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly*, *67*(268), 534-551.

Madva, A. (2019). Equal Rights for Zombies?: Phenomenal Consciousness and Responsible Agency. *Journal of Consciousness Studies*, *26*(5-6), 117-140.

McKenna, M. (2008). Putting the lie on the control condition for moral responsibility. *Philosophical Studies*, *139*(1), 29-37.

Mele, A. R. (2006). *Free will and Luck*. Oxford University Press.

Moran, R. (2002). Frankfurt on Identification. In Buss, Sarah & Overton, Lee (eds.). *Contours of Agency: Essays for Harry Frankfurt* (pp. 189-217). MIT Press.

Moran, R. (2012). Authority and estrangement. *Princeton University Press.*

Nagel, T. (1979). *Moral Luck*. In Nagel, T., *Mortal Questions*. Cambridge: Cambridge University Press, 24-38.

O'connor, T. (2007). Is it all just a matter of luck?, *Philosophical Explorations, 10:2,* 157-161.

O'connor, T. (2011). Agent-causal theories of freedom. *The Oxford handbook of free will,* 309-328.

Pérez de Calleja, M. (2014). Cross-world luck at the time of decision is a problem for compatibilists as well. *Philosophical Explorations*, *17*(2), 112-125.

Posner, M. I., & Snyder, C. R. R. (2004). *Attention and Cognitive Control.* In D. A. Balota & E. J. Marsh (Eds.), *Key readings in cognition. Cognitive psychology: Key readings* (p. 205–223). Psychology Press.

Pritchard, D. (2005). *Epistemic luck*. Clarendon Press.

Rescher, N. (2019). The Probability Account of Luck. In *The Routledge Handbook of the Philosophy and Psychology of Luck* (pp. 136-147). Routledge.

Pritchard, D. (2014). The modal account of luck. *Metaphilosophy*, *45*(4-5), 594-619.

Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127-190.

Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, *138*(3), 367-392.

Smith, A. M. (2015). Responsibility as answerability. *Inquiry*, *58*(2), 99-126.

Talbert, M. (2016). *Moral responsibility: an introduction*. John Wiley & Sons.

Teachman, B. A., Stefanucci, J. K., Clerkin, E. M., Cody, M. W., & Proffitt, D. R. (2008). A new mode of fear expression: Perceptual bias in height fear. *Emotion*, *8*(2), 296–301.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, M. (2012). Why the luck problem isn't. *Philosophical Issues*, *22*, 419-436.

Washington, N., & Kelly, D. (2016). Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227-248.

Williams, B. (1981). *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.

Yovel, I., & Friedman, A. (2013). Bridging the gap between explicit and implicit measurement of personality: The questionnaire-based implicit association test. *Personality and Individual Differences*, *54*(1), 76-80.

Zimmerman, Michael J. (1988). An Essay on Moral Responsibility. *Rowman & Littlefield.*

Article 3:
Resultant Luck and Many-Shots Actions

Abstract

Resultant moral luck suggests that actions are prone to luck in their outcome. According to an influential response, this indicates that one is not morally responsible for actions and their consequences. Thus, theories of moral responsibility are forced to adopt some kind of internalism, or resort to scepticism. In this paper, I argue that prominent cases of resultant moral luck are founded on the assumption that actions are events. I propose an alternative ontology of action as an ongoing goal-directed process with a "many-shots" structure that extends over time. Described this way, cases of resultant luck are not representative of ordinary action, and therefore, not generalisable. This diminishes the threat that resultant luck poses for theories of moral responsibility significantly. The proposal of action as a *many-shots* process is consistent with predictive coding, a plausible cognitive architecture which centres around error minimisation. Under this framework, cases of resultant more luck are no longer failures of action, but rather anticipated errors to be minimised within the ordinary process of action.

Section 1: Introduction

In this paper, I argue against the claim that because of apparent cases of resultant moral luck, people are not morally responsible for actions and their consequences. If control is a necessary condition for moral responsibility, one cannot be responsible for what is beyond one's control. Luck refers to factors that beyond one's control. Moral luck in particular refers to factors beyond one's control that appear to make a moral difference. This suggests a conflict. How can luck make a difference in one's degree of moral responsibility when it is beyond one's control? Either control is not a necessary condition for moral responsibility, or luck only appears to make a moral difference. I provide support for the second option.

The paper is a contribution to the growing literature on the possibility of a naturalistic account of free action in light of serious challenges, one of which is luck. I am concerned in particular with one kind of luck, namely resultant moral luck. I do not provide an argument against resultant luck in

general. Rather, I object to a certain use of apparent cases of resultant luck, which have been used to claim that people are not morally responsible for their actions and their consequences. Thus, resultant moral luck seems to force a choice upon us: One must either adopt a sceptical position or restrict the scope of moral responsibility to the internal mental life of the agent. It is exactly to this use of the argument that I object.

I begin in Section 2 by introducing moral luck in general, before I move on to a discussion of prominent responses. As most available accounts demonstrate, the task of explaining away the influence of resultant luck on an agent's blameworthiness is a hard one. In fact, after considering typical cases of resultant moral luck, one is pushed to two unattractive alternatives: internalism or scepticism. However, the cases themselves are built on an assumption that is rarely made explicit, namely that actions are events. I discuss the implications of that assumption in section 3 and argue that it might not be the best way to capture ordinary action.

In Section 4, I provide an alternative ontology of action, one that departs from the "one-shot" event structure and moves towards describing action as an ongoing goal-directed process with a many-shots structure. The process ontology suggests that actions are gradually unfolding over time, subject to change, and best captured in progressive sentences. Described under the alternative ontology, instances of resultant moral luck appear as special cases, no longer representative of ordinary action. While this does not exclude the possibility that luck may still influence an agent's blameworthiness, it significantly diminishes the force of the challenge by resultant moral luck. Thus, the credibility resultant moral luck provides to the view that people are not morally responsible for their actions is also diminished.

Finally, in Section 5 I examine whether there is a good candidate for a cognitive framework which can provide support for the process ontology of action. I turn to predictive coding as a plausible, naturalistic proposal for human cognition. I argue that predictive coding is consistent with the *many-shots* structure of action. Within a predictive coding framework, cognition is primarily engaged in error minimisation. Errors like the ones captured in cases of resultant luck are not necessarily failures, but rather integral elements of the many-shots process of action. Thus, predictive coding provides support to the claim that cases of resultant moral luck do not show that people are not morally responsible for their actions and their consequences. I finish with a discussion of potential objections, before considering a proposal for how to understand apparent cases of moral luck.

Section 2: Resultant Luck

Moral luck is luck which affects the moral evaluation of agents. The term *moral luck* was introduced in the contemporary moral responsibility debate by Bernard Williams (1981) and Thomas Nagel (1979). Both captured the idea that factors external to the agent may sometimes influence how blameworthy or praiseworthy one is. For instance, if an agent is more or less blameworthy by virtue of some chance occurrence, then that agent is morally lucky. Moral luck can affect the outcome of an agent's action, her disposition or character, the circumstances in which she finds herself, or simply the way the world causally determined.

A prominent type of moral luck is resultant luck. The term was coined by Zimmerman (1987), but the idea features in Nagel's original taxonomy too. Resultant luck refers to luck in the way that things turn out. In cases of resultant luck, we are inclined to morally assess people differently for an action, even though the outcome of that action was not something they could control. Agents under the influence of resultant moral luck are assessed as more or less blameworthy than identical agents with identical intentions.

A couple of clarificatory notes. In this paper, I refer to both *resultant luck* and *resultant moral luck*. Generally speaking, whenever resultant luck affects the moral evaluation of an agent, it is a case of resultant moral luck. It is a relatively uncontroversial statement that resultant luck exists. At least sometimes, it is not up to us how things turn out. Luck can affect the outcome of our actions in ways that we cannot foresee, control, or influence. However, that such luck presents a problem for theories of moral responsibility is more controversial. After all, it may be the case that while luck affects the outcomes of one's actions, it does not affect how blameworthy or praiseworthy we consider them for that action. Perhaps resultant luck is akin to gravity: it affects how things turn out, but there is nothing we can do about it and, generally speaking, we do not change our moral assessment of people because of its influence.

On a second clarification, the focus here is primarily on how resultant luck affects blame and blameworthiness. To be sure, praise and praiseworthiness are no less interesting. However, the possibility that one may be worthy of more blame because of some stroke of luck is particularly disconcerting for our ordinary understanding of morality. Correspondingly, discussions of resultant moral luck typically pay more attention to the blame side of moral responsibility, and so do I. Standard cases of moral luck readily demonstrate that luck often appears to make a difference to the

moral evaluation of agents. Take arguably the most common such case, which features in most relevant discussions since Nagel's essay (Nagel, 1979: 25): *drunk driving*.

> *Drunk driving*: Two drivers willingly and freely decide to drink and then drive. Everything about the two drivers and the external world is identical, except one fact: one drives home safely while the other, by stroke of luck alone, runs over and kills a pedestrian. Had the pedestrian walked off the pavement a moment sooner or later, she would not have been run over, and the second driver would have driven home safely, same as the first one.

Evidently, both drivers are equally blameworthy for drinking and driving. Yet it seems natural to blame the second driver for something else, and presumably more substantial: running over and killing the pedestrian. However, the two drivers are identical, and they act in exactly the same way. Up until the point where the second driver runs over the pedestrian, there is no difference between them. Presumably, their mentality is also identical. They share the same beliefs, desires, intentions, and so forth. All that is different about their respective outcomes is a matter of luck.

In *drunk driving*, luck seems to make a difference in blame between the two drivers. But should it? In other words, is the second driver more blameworthy than the first? If he is, it is because of the influence of luck in how things turned out. Essentially, this is the important question about luck in this case. By asserting that the second driver is more blameworthy than the first, one admits that luck may sometimes make a difference to the blameworthiness of agents. While this is a minority position, it has attracted some support (Hartman, 2017; 2020, Lang, 2021; see also Björnsson & Persson, 2012; Björnsson, 2017). In short, such views suggest that sometimes, resultant moral luck may affect the blameworthiness of agents.

However, many find the claim that resultant moral luck should influence moral assessment counterintuitive. We seem to have a strong natural intuition that moral assessment should be immune from luck. Philosophers have doubled down on that intuition, at least since Kant. Contemporary debate on moral luck from Nagel and Williams onwards, also centres on the same natural intuition, namely that luck should not affect how blameworthy or praiseworthy an agent is. In contrast, the view that resultant moral luck is consistent with moral responsibility suggests that differences in blameworthiness are warranted in cases where there is no difference between the agents' control or quality of wills. Since some kind of control or quality of will is typically accepted

as a necessary component of moral responsibility, I will set the view that accepts the influence of luck on blameworthiness aside for now.

Unless one is inclined to adopt the minority view that luck is consistent with moral responsibility, resultant luck should not make a difference in the blameworthiness of agents. Yet, this leaves the apparent differences in blame between the two drivers unaccounted for. In order to preserve core intuitions about morality and common moral practice, one is pushed to explain away the apparent differences in blame.

To do that, there are two primary options. First, one can argue that differences in blame between the two drivers are warranted, but they are not due to luck. Rather, there is a non-luck explanation of differences in *drunk driving* and similar resultant moral luck cases. Second, one can insist that there is no morally relevant difference between the two drivers. Proponents typically proceed by restricting the scope of moral responsibility, or by appeal to some independent constant factor which explains away the apparent differences in blameworthiness and blame alike. Generally, both options employ some kind of *denial strategy* (Nelkin, 2013), because they deny that luck should make a difference in how blameworthy an agent is.

Beginning with the first option, often best exemplified by the *epistemic argument* (Latus, 2000; a similar strategy is employed by Thomson, 1989). According to the epistemic argument, we are only inclined to blame the two drivers differently because we lack some epistemic knowledge about their actions (their exact intentions, desires, or plans). Had we access to this epistemic knowledge, we would still blame the two drivers differently not because of resultant moral luck, but because of the difference in epistemic attitudes between them. However, the epistemic argument soon comes into trouble. As cases like *drunk driving* demonstrate, it is possible to stipulate situations where the epistemic attitudes of the agents are identical, but moral assessments differ because of luck.

A prominent denial strategy of the second kind includes distinguishing between *degree* and *scope* of responsibility (Zimmerman, 1987; 2002). Zimmerman argues that degree and scope capture different aspects of responsibility. Degree expresses how morally responsible one is, while scope expresses what one is responsible for. According to this view, the two drivers are equally blameworthy in degree, but not in scope (since the second driver is also responsible for killing a pedestrian). Presumably there is some luck-independent constant factor which makes the two drivers equally blameworthy. But what could this factor be? It should be something common

between two agents, and still independent from luck. Since their respective outcomes differ so dramatically, that common factor can only be something internal to the agents, part of their mental states. I discuss the issues associated with restricting moral responsibility in more detail below. Another point against Zimmermann's account is that if the two drivers are equally blameworthy in degree but not in scope, it is ostensibly possible to hold people equally responsible even when they are responsible for different things (Khoury, 2018). Zimmermann's denial of resultant moral luck suggests that people can be morally responsible, even though there is nothing *for which* they are responsible, a claim that is counterintuitive, if not contrary to everyday moral practice. For present purposes, I rely on the assumption that Zimmermann's strategy does not adequately solve the issue.

Now to a more recent and promising resultant luck denial strategy, aimed at solving issues with existing ones. Consider two cases of an assassin named Lee, who finds himself under the apparent influence of resultant moral luck (Khoury, 2018: 1358-1359):

> *Assassin 1*: Lee is an assassin who has placed himself near a window on the fifth story of a building overlooking Elm Street. He carefully and skilfully draws a bead on his intended victim driving by below. He pulls the trigger, and the victim is shot and killed.
>
> *Assassin 2*: Lee is an assassin who has placed himself near a window on the fifth story of a building overlooking Elm Street. He carefully and skilfully draws a bead on his intended victim driving by below. He pulls the trigger but a bird flies into the path of the bullet and the victim is spared.

Before moving on, two clarificatory remarks. First, assassin 1 and assassin 2 are mentally identical up until and including the pulling of the trigger (ibid: 1359). In both cases, Lee has the same beliefs, values, and of course the same intention to kill his victim. Second, Lee meets the conditions of blameworthiness, whatever those may be. Whether that requires some kind of control, sensitivity to reasons, or some hierarchical alignment of his mental states, Lee is blameworthy.

Naturally, Lee seems blameworthy for killing his victim in assassin 1, but not in assassin 2 where the victim is spared. Since the flying of the bird into the path of the bullet is a chance occurrence, luck makes a morally significant difference between the two cases. That difference cannot be explained away by employing the epistemic argument, since we presumably have access

to all relevant facts about Lee's blameworthiness. Assassin is a compelling case of apparent resultant moral luck that cannot be explained by available denial strategies.

Ultimately, Khoury also denies that resultant moral luck should influence moral assessment, but his strategy is somewhat different from the ones previously mentioned. As a reaction to cases like assassin, which demonstrate how people's actions may be subject to luck, Khoury suggests restricting the object of moral responsibility to the internal mental life of the agent. This entails that people are only morally responsible for their willings (Khoury, 2018: 1367). Note that this brand of internalism about moral responsibility is different from Frankfurt's, who restricts moral responsibility to bodily movements (see Frankfurt, 1993). After all, one's bodily movement may be manipulated against one's will. An important upshot of such internalism is that it rejects moral responsibility for consequences. To say that an agent is responsible for the consequences of her action would entail admitting that resultant moral luck can make a difference in an agent's blameworthiness. Therefore, no one is ever blameworthy for the consequences of their actions, but only for their internal mental willings.

Taking stock, cases of resultant moral luck create a problem because they drive apparent intuitions about differences in blameworthiness, intuitions that are not easily accounted for. These cases drive a wedge between blameworthiness on the one hand and blame on the other. Intuitively, we want differences in blame practices to track differences in blameworthiness. The task is to explain away the morally relevant difference between agents under the influence of resultant luck. Strategies such as distinguishing between degree and scope of responsibility or the epistemic argument seem inadequate. A more recent such strategy involves adopting a kind of internalism, which restricts morally responsible action to a limited mental realm. If internalists are correct, we are only ever morally responsible for our mental willings.

Internalism may be a hard pill to swallow, but the alternative may be even worse. Denial strategies that aimed to warrant differences in blaming practices in some morally relevant factors unaffected by luck seem inadequate. In response, internalism argues that while the difference in blaming is not warranted, people may still be morally responsible, even if only for their willings. Given the assumption that denial strategies are problematic, if the internalist strategy fails, then one is pushed to embrace some sort of scepticism about the possibility of moral responsibility in general. If people are not even morally responsible for their internal mental states, is not moral responsibility simply impossible?

The debate on resultant moral luck seems to lead to a choice between internalism and scepticism about moral responsibility. One is pressed to either shrink the domain of moral responsibility to the mental, or to accept the influence of resultant moral luck and admit that some form of scepticism (or at least revisionism) about assigning blame and praise is in order. Both alternatives are unattractive. On the one hand, internalism suggests that people are not morally responsible for their actions and their consequences. On the other hand, scepticism involves discounting strong natural intuitions about moral responsibility.

Section 3: Resultant Luck and Event Ontology

With the available reactions to cases of resultant moral luck being unsatisfactory, it is hard to imagine what progress on the matter might resemble. However, a further option remains, namely to examine the structure and treatment of the cases themselves and consider how that might affect our conception of luck, and of free action in general. In the remainder of the section, I want to argue that much of the force of the argument for the existence of resultant moral luck depends on the way the relevant cases are structured.

Consider again the cases discussed above: *drunk driving* and *assassin*. Both describe agents acting in the world, and both involve outcomes that are undesirable, morally significant, and beyond the agents' control. The unlucky versions of both the drunk driver and the assassin end up in some unfortunate outcome, because of factors that are seemingly external to them. This is indicative of how resultant moral luck is introduced in the literature. Examples typically involve an agent who acts in a certain way, but under the influence of luck, the action ends up in some unforeseen outcome. Within this case structure the case for resultant moral luck seems convincing. Having acted in one way, the agent ends up in a situation because of factors outside her control. The outcome is the result of good or bad luck and blaming or praising her more than we would otherwise seems to suggest a conflict with common intuitions about moral assessment.

Now, I want to suggest that resultant moral luck cases work in three steps. First, actions such as driving drunk or attempting to assassinate someone are prone to luck regarding their outcome. This first premise is a simple statement of resultant luck. Second, actions in general have this structure. In fact, *drunk driving* and *assassin* look like pretty much ordinary examples of action. They are only special because luck *in fact* affects their outcome. Otherwise, nothing about them

seems unusual. From these two premises, a rather more substantial conclusion follows: all actions are subject to luck in the outcome. Consider this in argument form:

1. Actions like the ones described in *drunk driving* and *assassin* are subject to luck in how things turn out. (Resultant moral luck premise)
2. *Drunk driving* and *assassin* are examples of ordinary action, i.e., what is true of them is true of actions generally. (Generalisability premise)
3. Therefore, all actions are subject to luck in how things turn out.

While that argument may seem sweeping, much of its force lies on a single, rarely explicit assumption. This is the idea that actions are a specie of events. Conceiving of actions as events is far from unique to the moral luck debate: it is arguably the default ontology in philosophy of action at least since Davidson (1963; 1974). Understanding the way that we ordinarily talk about action has given rise to the problem of *variable polyadicity* (see for instance Ludwig, 2010). In its essence, the problem appears when one attempts to provide an adequate explanation of a fact about ordinary language, namely that from an utterance such as "Jones buttered the toast at midnight with a knife" one can infer several more limited claims like "Jones buttered the toast" or "Someone did some buttering at midnight", and so forth. Coming up with a satisfactory explanation has been one of the main tasks of traditional philosophy of action, and Davidson has provided the most conspicuous solution to the problem.

Davidson's solution suggests that actions are event particulars that can be described in many ways and warrant a specific inference pattern. This has two important implications. First, the same event can be given many action descriptions. For instance, "Jones buttered the toast" and "Jones made buttering noises" may be descriptions of the same action. Second, these descriptions of individual actions quantify over actions conceived as particulars. From hearing "Jones buttered the toast with a knife at midnight" one can infer that "Jones buttered the toast with a knife", that "Jones buttered the toast", and so forth.

All this suggests that actions are the kinds of things that happen at a specific time, in a specific way, under specific conditions. The most common way to understand this in metaphysical terms is to think of actions as events. However, this largely orthodox metaphysical position makes a crucial difference once it is applied to the resultant luck debate. Assuming that actions are events creates a barrier between the action on the one hand, and its outcome and consequences on the other. In cases

like *drunk driving* and *assassin*, that assumption invites conceiving of the action as one event, and its consequences as other particular events. Thus, a clear distinction is created between the action and all that follows it, including its outcome.

Now, it is worthwhile to flesh out the claim that resultant moral cases are premised on the assumption that actions are events. There are two distinct claims here. The first one is that the actions in cases like *drunk driving* and *assassin* are understood as events. Once the event conception of actions takes root, a distinction between actions and outcomes emerges, thus making the case for resultant moral luck not just plausible, but convincing. The second claim suggests that it is possible to generalise from specific cases to all actions if we conceive of them as events. This generality claim suggests that if actions are events, they share a structure by which the action itself is distinct from its consequences and outcomes. Under this description, all actions are subject to luck regarding their outcome.

To sum up, prominent cases such as d*runk driving* and *assassin* demonstrate that such actions are prone to moral luck in the outcome. But dialectically, they are used to argue for something more general. For instance, that we cannot be responsible for consequences, or that we are only responsible for our mental willings. These cases can only be used to make this general argument because they assume that all actions are events. So, rather than arguing that all actions are susceptible to luck in the outcome, these cases are more accurately used to argue that actions are subject to resultant luck if they are conceived of as events.

Before we can evaluate that assumption, some more details are needed on what it means to conceive of actions are events. For this, consider again the cases of d*runk driving* and *assassin*. Several features stand out. First, the actions in both cases are *spatiotemporally located particulars*. As mentioned above, actions exist in the world under specific conditions. In orthodox philosophy of action, an action is a particular event, specifically located in time and space. The driver drinks, then drives, then runs over the pedestrian. The assassin shoots, then misses.

Second, such events possess a *completeness* feature. This is easily demonstrated by the fact that they are expressed in past aspects. Often, this is masked by the historical present (i.e., the assassin shoots, the drunk drinks, then drives). But the actions to which they refer lay wholly in the past. It is commonplace in contemporary action theory to speak of actions in the past tense, as Davidson also does. Yet, it is far from obvious that this is the best way to conceive of action. Many

have followed this vein of objection, perhaps most famously Thompson's naive action theory (Thompson, 2008).

The third relevant feature is that events are *up to nature*. If actions are events, they are part of the event causal structure of the world, and thus, their occurrence and subsequent consequences are at least partly out of the hands of the agent. This feature raises familiar questions about how events cause each other and what counts as a basic action. Consider for instance the Shakespearean tale of Claudius poisoning the king by pouring poison into his ear. When exactly did Claudius kill the king? Was it when he poured the poison into his ear? Or when the poison affected the king's internal organs? Or was it when finally, the king succumbed and died? It is far from clear exactly when the Claudius' action was completed.

Fourth and final is the fact that events are *non-changeable*, even though they are themselves changes. The argument that events cannot change in respect to intrinsic properties is not new (see Dretske, 1967; Galton & Mizoguchi, 2009; Steward, 2012). That final feature will become important in the next section.

A potential objection at this point might question whether the event conception of action is necessary for actions to be prone to resultant luck. Suppose we reformulate our action ontology, so that actions are not understood as events, but as something else. Would that still make them subject to resultant luck? I think this is a worthwhile project, and I engage with it in the next section. Another potential objection may be directed at the generality claim. Do all actions need to be events, for resultant moral luck to generalise from specific cases to potentially all actions? Here the answer is no. Perhaps there is another way to generalise. What I have argued is that assuming that actions are events provides at least one way to generalise resultant moral luck cases to all actions.

Now we can consider why one might think the assumption that actions are events false. What is the problem with thinking of actions as events? If events have the static nature of completed particulars, then describing certain garden-variety actions as events seems inadequate (see Hornsby, 2012; and Steward, 2012 for detailed arguments). For instance, an event ontology of action seems insufficient to account for an action that changes over time (if, say, it becomes more intense) and yet remains naturally understood as a single action. As Steward puts it "we engage in changings and not in changes, movings and not in completed motions. The things we engage in are things which are themselves susceptible of change -and this means they cannot be events." (2012: 384).

Additionally, conceiving of actions as event particulars gets little traction in explaining how an agent contributes to her action, thus allowing for the issue of the disappearing agent to arise (Hornsby, 2012). However, this is not the place to engage with that issue.

Where does this leave us? The argument for resultant moral luck generalises from specific cases to actions in general because it assumes that all actions are events. However, there is some indication that at least some actions are not captured by this ontology of spatiotemporally located completed particulars, which cannot change themselves. At least some actions seem to be changeable or variable, and they seem to have a more continuous or unbroken profile over time. So, if resultant moral luck depends on the assumption that actions are events, but there is reason to think otherwise, it raises the question: what if actions are not events, but something else? Would actions differently conceived be subject to resultant moral luck? Or would it, more likely, be more of a case-by-case assessment? I discuss the plausibility of a different ontology of action and the implications for the debate on resultant luck in the next section.

Section 4: Action as a Many-Shots Process

Prominent arguments claiming that luck makes no difference to blameworthiness seem unable to adequately explain apparent cases of resultant moral luck. In reaction, recent strategies restrict the domain of morally responsible agency to the internal mental life of the agent. However, classic cases of resultant moral luck rely on a rarely explicit assumption that actions are events. In what follows, I consider what an alternative ontology of action might look like, and how it might affect the problem of resultant moral luck. I argue for describing actions as ongoing processes of goal-directed behaviour, better captured by a *many-shots* structure. Following that conception, typical cases of resultant luck are far from representative of ordinary action, but rather special occurrences that cannot be used to make general claims about action more broadly.

Before discussing an alternative action ontology, two things should be noted. First, suggesting that actions are not events does not mean that they are necessarily protected from the influence of luck. Perhaps actions are something other than events, but whatever they are is still subject to luck. Or maybe some actions may still be events, and they would be subject to luck. For the alternative ontology presented here, I rely on accounts whose basic tenet is that actions are not events. However, it is not important to my argument that no action is an event. Some actions could still be events under some description. However, even though one might be able to describe actions in an

event-like fashion after being completed, I argue that is overall better to think of actions as processes, not events.

Second, while thinking of actions as events is a standard position within philosophy of action, it does not need to be so. Davidson's solution to the problem of variable polyadicity does not depend on actions specifically being events. This point has recently been made by Sarah Paul (Paul, 2020: ch.4). The Davidsonian account depends on interpreting the semantics of action descriptions as quantifying over action particulars, in a way that would allow inference patterns such as inferring "there was at least one buttering" when reading "Jones buttered the toast angrily at midnight". One way to do that is to think of actions as events, but at least in principle, this conclusion is not inescapable. One could have different ontological conceptions of the action particulars. Therefore, an alternative to the event ontology of action may still be within the framework.

With these points clear, we can ask what actions might be if they are not events. There are at least two reasonably developed alternatives to the traditional ontology of action. Instead of thinking of actions as events, one can think of them as *processes* or as *stretches of activity*. While these two accounts are not explicitly meant to complement each other, they present a largely compatible alternative to the orthodox view. First, let us consider what it means to conceive of actions as processes. Processes, unlike events, can change. In fact, on a common understanding of processes, they are changings (Steward, 2012). Their constitutive parts -that is, whatever makes them up- can themselves change. And the rate of change by which these constitutive parts can change, can vary itself. Change may happen in a way that is smooth, abrupt, or momentary.

Along similar lines, one can think of action not as a category of particulars, such as events, but as *activity* (Hornsby, 2012: 233). Hornsby's account attempts to tackle the problem of the disappearing agent in explanations of action by suggesting that when an agent acts, she "engages in activity, where no activity is any particular" (ibid: 234). Conceiving of actions as activity conveys an interesting analogy between time and space. Activity makes up the temporal world, much like bits make up the physical world. Just as space consists of physical things, time consists of activity. Rather than any event particular, activity is the basic ontological category which explains action.

Summing up the alternative ontology of action. There is a way of conceiving actions not as events but as unfolding processes of activity. The ontology of action as a process describes the agent's journey towards a goal. Several aspects of that journey are susceptible to change, as one

would expect of a goal-directed process that unfolds gradually over time. Here we see a departure from the traditional view of action, which allows for an understanding of action not as a series of completed events, into a more continuous and less temporally defined affair, changing over time.

Having sketched the case for an alternative ontology of action, the next step is to consider exactly how it departs from the traditional event ontology. In the previous section, I presented and briefly discussed four features of events. If actions were events, then these features would also characterise actions. But if actions are rather not events, it is worthwhile to consider how the same features might differ. Two of these features are the most telling: *completeness* and *non-changeability*. In what follows, I argue that under the proposed alternative ontology, actions often do not have these features.

Begin with completeness. If actions are ongoing goal-directed processes, they are better captured as extended over time rather than completed particulars. And action is perhaps best expressed using progressive aspects, such as the present progressive or the present perfect progressive, when the action lies in the present tense. After all, we rarely know when or if the action will be completed. Say I decide to get up from the couch and turn on the light. If, as I get up and start walking to the switch, someone asks me what I am doing, it seems natural to reply "I am turning on the light", even though I nowhere near the switch. If I flip the switch and nothing happens, I may still be turning on the light, as long as I maintain that goal. My objective is to have the light turned on, but even if I never do that, I could still reasonably claim that I was turning on the light. Even when the action lies in the past, aspects such as the past progressive or the past perfect progressive seem to better capture the variable nature of action. Perhaps as I was walking to the light switch, I stopped to drink some water. Of course, there may be cases where the action is irrevocably completed. Most of the time, however, we describe actions that stretch over time, towards a goal. Open progressive sentences indicate progress towards some goal (Wolsfon, 2012). Such sentences are arguably better at capturing a structure that is common to all actions. In any case, according to the proposed alternative ontology of action as a process, completeness is reversed into openness.

Next, consider the non-changeability feature. Events are themselves changes, but they cannot change in respect to their intrinsic properties. A buttering at midnight is a change, but it does not change itself. So, if actions are events, they are unchangeable. In contrast, if actions are ongoing processes they can not only change, but the rate of change can itself vary over time. In fact, it is

possible to say that an action begins being fast, then becomes slower, or angrier, or shy, and still be describing the same action as it evolves over time. As soon as one departs from the traditional event ontology, action emerges as a phenomenon much more capable of change than originally thought.

There are reasons to think that ordinary action is best captured by progressive aspects. In contrast to the perfect, sentences in the progressive demonstrate a kind of progress towards something, typically towards a goal (Wolfson, 2012). Progressive sentences are characterised by temporal openness, which captures the fact that actions can be paused and resumed, and the all the while be ongoing. Furthermore, progressive sentences are good at capturing the so-called imperfective paradox, the fact that for an action to be ongoing in the present, it does not need to be completed in the future. While linguistic in nature, these considerations about progressive sentences give support to the idea that actions are more like ongoing processes and less like completed events. That said, it may be the case that we often use simple present or perfect aspects to describe action. However, progressive sentences are more precise at capturing ordinary action.

Under this alternative ontology, actions are no longer captured by appeals to events and completed particulars. The process ontology presented here suggests that actions are goal-directed processes, ongoing, and best captured by progressive sentences. Actions can change over time, they can be paused and resumed, and still be actions. Far from the fixed, one-shot structure exhibited by the event ontology, actions are processes that often require many shots. To complete an action, an agent may need more than one tries. She may try once, fail, try again, pause, resume, try again, and yet be engaged in the same action throughout. To understand the relevance of this for the debate on moral luck, we need to revisit the two familiar cases of apparent resultant moral luck, *drunk driving,* and *assassin*. If actions are ongoing processes of goal-directed behaviour, *drunk driving* and *assassin* no longer resemble cases of everyday, ordinary action that are generalisable.

Begin with *drunk driving.* Under the process ontology of action, *drunk driving* is a special case, and hardly generalisable. A harrowing case, to be sure, but an exception nonetheless. Calling it special does not mean it is not realistic or even likely to happen. It means, however, that it is not representative of the structure of ordinary action. Importantly, drunk driving is not special because the case is based on the assumption that actions are events, but for two different reasons. First, it is special because it is a case of recklessness, where we are dealing with predictable unintended consequences. While there are interesting things to say about such cases, they are hardly representative of everyday action. Second, drunk driving is special because it has a particular "one-

shot" structure. It requires special constraints to make an action a one-shot case. The agent acts, and then has no chance to try again, to rectify or to change her conduct in any way. What is done is done. Once the pedestrian is knocked over, the situation is final and irreversible.

But ordinary action is rarely like that. Normal cases of action have a *many-shots* structure. Consider again the case of Lee the assassin. He shoots, misses, and the action is over. While that may happen, it would not represent what we usually think an assassin would do. Indeed, if Lee is a professional, we should expect him to try and shoot his victim again and again. Imagine you have hired Lee. You would expect him to finish the job. If for whatever reason he only tries once (say he only had one bullet, or he was worried he would get spotted, or he needed to leave to fulfil another murder contract), you might blame him for giving into a flawed character trait which makes him bad at being an assassin. Naturally, we expect Lee to do all within his power to shoot and kill his victim. If he misses the first time, he should try again. If he only has one shot, the first question is: why does he not have more shots? And finally, if indeed he can only try to kill his victim once and he fails because of luck, it is important to note that this is also a special case, far from representative of ordinary action. Assassin has a clear one-shot structure, which does not represent ordinary action. Therefore, it does not generalise.

So, under the proposed alternative ontology according to which actions are ongoing process of goal-directed behaviour usually involving many shots, it is no longer possible to generalise from cases like drunk driving and assassin to all actions, because such cases do not share the common structure of action. Ordinary actions often stretch over time and allow for a *many-shots* structure.

Consider *Adonis pouring wine*:

> Adonis tries to pour a cup of wine while he is lying on a couch. As he is pouring, he spills most of the wine on the floor. He then wipes the wasted wine with a cloth and tries to pour himself a cup once more. Then he discovers there is no more wine left in the crater. Adonis gets up, walks to the cellar, and fills the crater with fresh wine. Finally, he fills his cup with wine and drinks contentedly.

It seems to me that *Adonis pouring wine*, while admittedly comical, is a case of ordinary action through and through. It is not special in any meaningful way. Adonis has a goal, which he tries to accomplish by the means at his disposal. He tries, fails, tries again. He adapts and reacts to

unforeseen obstacles that appear along the way. Finally, after multiple tries and perhaps more effort than he initially thought would be necessary, he achieves his goal: a cup full of wine.

Imagine Adonis giving up on his intention to pour himself some wine once he discovers his wine was spilled, or after he cleans the wine from the floor, or even after he realises the crate is empty. That would make very little sense given Adonis' desire for a cup of wine. Desires that function as intentions which guide an agent's action should presumably persist in the face of obstacles. Otherwise, intentional or rational action would be too unstable. Indeed, every day action is stable. Even where there are hiccups along the way, one proceeds to try again until the goal is reached.

Actions of this sort are much less likely to be subject to luck in the outcome, at least not in the way suggested by apparent cases of resultant moral luck. In the previous sections, the argument for moral luck was based on thinking of actions as events. According to the alternative ontology presented here, action is an ongoing goal-directed process with a *many-shots* structure. Is this an argument against resultant moral luck? Or in other words, does it provide reason to think that people are not more or less blameworthy because of resultant luck? Not quite. But the force of the cases, as generalisable instances of ordinary action is diminished. With them, the force of argument for moral luck is also diminished. In the next section, I present a prominent naturalistic cognitive framework that may underlie this alternative ontology of action.

Section 5: A Cognitive Architecture for Action

5.1 Predictive Coding

So far, I have proposed an alternative to the event ontology of action, which seems to avoid some of the pitfalls of the one-shot structure of action. In particular, the conception of action as an unfolding, goal-directed process involving many shots diminishes the generalisability of prominent cases of resultant moral luck, because such cases are no longer representative examples of everyday action.

One concern with the alternative process ontology of action is that it does not align with an otherwise naturalistic view about action, because it does not lie on empirically sound ground. Often, proponents of views like the one I have presented rely primarily on linguistic evidence (of the sort presented by Wolfson, 2012) or provide a non-naturalistic explanation of action (such as

Thompson, 2004). Such views claim that human cognition is normatively unique, so that it is best not understood and investigated within the empirical model of the natural sciences. Furthermore, naturalistic causal theories of action often explicitly endorse an event-causal metaphysical framework.

However, my aim is different. I want to demonstrate that the *many-shots* process ontology of action is consistent with a naturalistic approach to action and human cognition in general. To do that, in this section I present an empirically well-supported framework for a cognitive architecture that supports the idea of action being an unfolding goal-directed process. That process is naturally approached with a many-shots structure, in contrast to the one-shot structure of traditional event ontology. That framework is predictive coding (Rao & Ballard, 1999; Huang & Rao, 2011; Friston, 2009; 2010; 2018), also known as predictive processing (Clark, 2013; Hohwy, 2013; Colombo, 2017). For the rest of the paper, I will use the term predictive coding.

Predictive coding forms the core of a broad and ongoing research programme. As such, empirical evidence for it is qualified and even contentious. However, there is good reason to think that this proposal can provide answers about human cognition that competing frameworks cannot, and that it captures something true about human cognition. Furthermore, only recently have philosophers started to engage with this otherwise prevalent programme to inform long-standing debates in philosophy of mind and action (see for instance Hohwy, 2020). Predictive coding refers to a cognitive architecture which employs *prediction* and *prediction error minimisation* as the primary task of a mind (Clark, 2013; Hohwy, 2013). According to Andy Clark, the human mind "is basically a prediction machine" (Clark, 2011). Predictions guide perception, thought, and action by anticipating and explaining sensory and sensorimotor input. This is a unified process which underlies all cognitive activity. Minimising the difference between the agent's predictive model and the various sensory information which the agent encounters is the *only* task of the human mind.

According to predictive coding, the nature of cognition is probabilistic. In order to make predictions about the current state of the world and the likelihood of different events occurring, the mind uses already existing knowledge about the external world and previous predictions to calculate a wealth of subjective probability distributions. These probabilities capture the degree of confidence assigned to each possible outcome, and they are constantly updated until the moment the agent is required to choose one outcome in action. Predictive coding is not alone in suggesting that cognition is probabilistic. For instance, Bayesian-brain hypotheses (see Hoyer and Hyvärinen, 2003;

Sanborn and Chater, 2017; also, Friston, 2012), claim that subjective probability distributions are updated following Bayes' rule. These hypotheses typically adopt either a sampling or a variational method. Predictive coding uses a variational method to approximate Bayesian inference.

Another important aspect of predictive coding is the centrality of error signals. Identifying, incorporating, and minimising prediction error signals is the first step in inferring the states of the world (Friston, 2010). When there are discrepancies between the internal predictive model of the agent on the one hand, and sensory input from the world on the other, error signals are generated. The sum of all error signals is typically called *free energy* (Friston and Stephan, 2007; Friston, 2009, Hohwy 2021). It is the primary task of the mind to minimise free energy. That involves either updating the model or engaging in active inference by sampling new evidence from the world to confirm the existing model. In the first case, the subjective probabilities of the model are modified to account for the detected error signal and select between several possible models, thus minimising free energy. In the second case, the cognitive capacity of the agent is employed to gather new information from the world, thus driving exploration of the environment. In both cases, the same fundamental process is demonstrated, namely predictive error minimisation. To be sure, predictive coding is not the only framework that employs error minimisation. Most computational models in machine learning and statistics also do (see for instance Bishop, 2006; Barber, 2012). The difference is that according to predictive coding, predictive error minimisation is not just one goal, but the *only* goal of human cognition.

5.2. Bayesian Update and Active Inference

Now let us take stock. Predictive coding suggests that the at the heart of human cognition lies a generative predictive model of incoming sensory information from the world. Prediction mismatches between that model and the world create error signals, which are called free energy. Free energy is minimised in two ways: either by Bayesian update of the current model to fit the sensory input (as in ordinary perceptual learning), or by engaging in active inference towards confirming the existing model (by taking appropriate action, for instance sampling new sensory evidence).

To see how all that may work in practice, consider a case where a person is placed blindfolded in a room. She is confused and naturally tries to make sense of her situation and her environment. While her visual input is inhibited, she can still touch, smell, hear, and even taste

things around her. We can reasonably expect the blindfolded agent to make various guesses about her surroundings. In fact, she will probably soon have a whole array of potential explanations about what is happening in the room. Each of these explanations will in turn be assigned different priors. It is conceivable that the guess "I am blindfolded in a room" has a much higher prior than the guess "the sun stopped shining". As she engages with her environment, some guesses are bound to be wrong, but some may be confirmed. Say she picks up a roughly round object, which feels and smells like an apple. When she bites into it, it also tastes like an apple. Reasonably, she might confirm her guess and add the apple into her partial mental image of the room. Other guesses maybe be harder to confirm. Depending on the strength of the prior, the agent might stick to her existing model and discard evidence to the contrary. For instance, if the probability value of "I am sitting on a chair" might be high enough for the agent to ignore the fact that nothing about where she is sitting resembles a chair. Instead of discarding the "I am sitting on a chair" model, she might keep sampling for new evidence that explains away the data until the model is confirmed.

As the case above demonstrates, what explains away the sensory data is not just the probabilities, but also the content of the model. Changing the content of model is akin to adopting a new one. All the while, the agent has simultaneous competing models, which all undergo Bayesian updating. And while the agent will always choose the model with the highest probability, the way this happens is not straightforward. For instance, if there is a model with a very strong prior and some conflicting evidence appears, the agent might be reluctant to adjust the probability of the prior. So, she might sample new evidence in an explorative way to confirm her existing prior. Importantly, both adjusting the prior because of sensory input and sampling in an explorative way are part of a process that takes time. Even in cases where there is only one model and only one explanation, confirming that explanation is still a process. When the agent engages in active inference, the goal state comes in the form of sensory predictions. By error minimisation, the agent changes the world in order to fit the goal.

5.3 Predictive Control for Action

With the predictive error minimisation process laid out, we can move from cognitive architecture and active inference to human action. One way to do that is via control. According to an influential view proposed by Daniel Dennett (1984), agents exercise control to arrange the world in a particular way. The role of control is to minimise the difference between a goal state on the one hand, and the sensory input about the state of the world on the other. The goal state is of course

internal to the agent, a model to which the state of world is compared. Predictive error minimisation is the process by which the agent conforms her predictive model to the world. Control is the process by which the agent conforms the world to her model of it.

Such a process is not without hiccups. Agents navigate a world that is uncertain and risky. Information is patchy and sometimes untrustworthy, the environment changes constantly and sometimes in unpredictable ways. It is exactly this incomplete and uncertain informational environment that feeds into deliberation, thinking, and action. Given that, the degree of control one has over both the environment and over her own conduct is contingent on one's partial view of an ever-changing world. Within this informationally patchy and changing environment, minimising predictive error is a process. One works progressively towards filling the gaps, making the picture of the world clearer. But at the same time, that picture is itself moving. The world changes. So, there will always be errors. Even if the agent becomes exceptionally good at predicting and explaining sensory input from the world, the world itself will change, rendering even the best predictions inaccurate. As a reaction, the agent will then engage in model update or active inference.

Consider again the example of *Adonis pouring wine:*

> Adonis tries to pour a cup of wine while he is lying on a couch. As he is pouring, he spills most of the wine on the floor. He then wipes the wasted wine with a cloth and tries to pour himself a cup once more. Then he discovers there is no more wine left in the crater. Adonis gets up, walks to the cellar, and fills the crater with fresh wine. Finally, he fills his cup with wine and drinks contentedly.

Now we can examine *Adonis pouring wine* as an example of active inference in terms of action control. Adonis has a predictive goal state, namely a full cup of wine. He predicts grasping a cup, and pouring wine from the crater into his cup. This is all cashed out in the form of different sensory predictions. He predicts certain tactile inputs, certain activations of joints and muscles, and so forth. But alas, he is wrong. The wine ends up on the floor, not into his cup. Some of his predictions were obviously wrong. So, what does he do? He engages in active inference. He moves in such a way so that he can turn his prediction of a full cup of wine true. Adonis' goal takes the form of different sensory predictions. To achieve his goal, Adonis engages in active inference, changing the world to minimise the error in his model. This is exercising control.

Active inference, as demonstrated briefly in the case above, is obviously a many-shots process. To accurately predict, explain, and manipulate the world around us, we need to try and fail multiple times. The sort of control that is required for action can be understood in these terms. Through a process of active inference, the agent is involved in a many-shots process of exercising control, by minimising the error of her predictive model of the world. As her model becomes more accurate, she accrues a higher degree of control over her interaction with the environment. One such interaction is what we ordinarily call action.

In no way is this discussion meant to be proof that predictive coding is the correct model for deliberation, decision-making, and action. Rather, the goal is to demonstrate that predictive coding aligns well with the process ontology of action, which in turn has important implication for the discussion on resultant luck and moral responsibility in general. Predictive coding suggests that the fundamental function of the mind is predictive error minimisation. Control for action within this general framework is characterised by a many-shots structure. This provides way to couple the notion of active inference with the idea of control for action. When engaging in active inference, the agent changes the world until it fits her goal, as captured by her predictive model. In terms of action control, this means that the agent makes a prediction about some proprioceptive input. When she gets a contrary input, she can change that input by moving her body in different ways. Thus, she exercises control in different ways, in order to conform the world to her model. So, the picture of the mind as presented by predictive coding aligns with a view of action as a many-shots unfolding goal-directed process.

5.4 Objections

Before closing this section, I consider two objections to the claim that the predictive coding framework is consistent with the process ontology of action. First, one might wonder whether the two accounts are referring to the same type of process when they discuss action, or if they are in fact talking past each other. Given that predictive coding is a way to model cognitive processing, it may be that within the framework, the way to understand action is different from the common understanding of action in philosophy. But that would be problematic. That is, only if action as discussed in the predictive coding framework is the same type of process as discussed in philosophy of action and moral responsibility would the consistency between the process ontology of action and the predictive mind hypothesis be significant. If the two are talking about different processes, it raises the question of whether any consistency between them is at all informative in the first place.

This first objection can be understood in terms of distinguishing between *personal* and *subpersonal* levels of explanation (Dennett, 1969, for a more recent defence see Hornsby, 2000). In philosophy of action, the focus ostensibly is on the personal level. We are interested in an explanation of action that retains a sufficient (causal or otherwise) role for the agent. If the necessary and sufficient conditions for something to count as an action are best spelled out in subpersonal mechanisms, then it is unclear exactly what contribution the agent's personal level states might play. And if the agent's personal level states do not contribute to something being an action, then it is hard to see how such a thing is something the agent does in the first place. So, according to the personal/subpersonal distinction, it is important that we can explain an action by appeal to the agent's personal level states. When it comes to predictive coding, things seem to be different. At its core, the framework is meant to model subpersonal cognitive activity. The problem therefore is this: Philosophers and predictive coding theorists seem to appeal to different levels of explanation, and therefore different concepts of action (see Holton, 2016 for more on this point).

Now to the reply. While it is true that the personal-subpersonal distinction has been influential in philosophy of action, and while it is true that predictive coding theorists make little reference to explicitly *personal level* cognition, it is not true that we are dealing with different concepts of action in the two discussions. Once one adopts a predictive coding approach to explain and model cognition, the distinction between personal and subpersonal levels of explanation is drawn into question. In fact, many prominent theorists both within and outside the predictive coding paradigm deny the distinction in its original form altogether (Clark, 1989: ch.3; Bermudez, 2000; for an overview see Colombo, 2013). Now, it is beyond the scope of the paper to argue for or against the personal/subpersonal distinction. That said, predictive coding is a framework that is meant to work on all levels of cognition. Whether the framework adequately captures cognition is a different question. But if it does, then it explains the sort of action usually discussed in philosophy, and therefore, it can inform philosophical discussions about action and moral responsibility.

The second objection can be formulated in the following way: if predictive coding is true, then all motivational states are predictions. Motivational states then are best understood as beliefs about the world. But if all motivational states are beliefs, then most leading theories about motivation, including the Humean theory of motivation and Bratman's theory of intention, are wrong.

Begin with Humean theories of motivation. According to a Humean theory, beliefs and desires are distinct states, with distinct contributions to motivation. Belief is not sufficient for motivation - desire is also required. However, according to predictive coding all motivational states are predictions about the world, and they are spelled out in degrees of belief. Thus, if beliefs and desires are both degrees of belief about the world, they cannot be distinct states. Therefore, the Humean theory of motivation must be wrong.

Now to Bratman's theory of intention. According to Bratman (1987), intentions are a sui-generis mental state, distinct from beliefs and desires, which support our practical reasoning and action, and feature heavily in planning. Similarly here, if intentions are best understood as degrees of belief about the world and not distinct from other mental states, Bratman's conception of intentions must be wrong. Note that the buck does not stop with Bratman's theory. For instance, the desire theory of intention (see Ridge, 1998), which suggests that our strongest desires have the motivational profile of what Bratman calls intentions also runs into trouble with predictive coding. Any theory that conceives of desires, beliefs, or intentions as distinct will similarly be refuted by the truth of predictive coding.

The general issue, then, is that predictive coding reduces beliefs and desires to predictions, thus supporting a kind of strong cognitivism (see Colombo, 2017). According to strong cognitivism, intentions are a type of belief about the world (Marušić & Schwenkler, 2018). When an agent intends to do something, she believes she will do it based on her practical reasoning. Furthermore, one may adopt corresponding view about desires, conceiving them as a sort of belief about the world (Lewis, 1988; 1996). While adequate explanation and defence of the belief theory of intention and the belief theory of desire are beyond the scope of this paper, it suffices to say that they demonstrate viable and philosophically respectable alternatives.

On the face of it, is true that predictive coding leads to a version of strong cognitivism about motivational states. For the sake of this argument, I apply this approach of predictive coding, and together with that I assume some version of strong cognitivism about intention and desire. Strong cognitivism is a substantive position to hold, but it is a prominent position, and it has received considerable support. Furthermore, the truth of strong cognitivism is independent from the truth of the predictive coding framework. Supporters of strong cognitivism consider it a viable position regardless of whether that position is corroborated by predictive coding or not. In any case, the truth of strong cognitivism is beyond my purpose. I merely argue that predictive coding presents a

plausible picture of cognition. Surely it comes with its own set of problems. But whichever are the best arguments that the strong cognitivists have developed for their position, I can adopt them in my reply to the second objection.

Section 6: Concluding Remarks

Before closing, I want to consider the relevance of the proposal presented here for the debate on luck and moral responsibility. I started the paper with a common challenge to the possibility of morally responsible free action: resultant moral luck. After looking at different possible replies to the challenge, I ended up arguing that the force of the challenge from resultant luck comes from the assumption that actions are events. Once we abandon this event ontology of action, the force of the luck challenge is diminished.

Consider again the proposed internalism about moral responsibility, according to which people are only ever responsible for their tryings or willings. Such internalism is motivated by cases like *assassin*. However, under the process ontology of action, cases like *assassin* seem much less daunting. Now with a plausible a cognitive framework which supports a many-shots structure of action as an unfolding process of minimising errors in the agent's predictive model of the world, what can be said about moral luck in general?

Regardless of the objective nature of the universe, the agent faces the world from a perspective with an unavoidable degree of uncertainty. Thought and action are not immune to this uncertainty. If predictive coding presents a plausible picture of the mind, this uncertainty is built into the way our minds operate. We are made so that minimising uncertainty about our environment is the main task of our mind. One may ask what implications this has for discussions of moral responsibility and luck. Given a predictive coding picture of cognition, we should consider the possibility that our cognitive psychology serves the purpose of getting luck under control. Action has a many-shots structure because control of action needs to get chance under control.

Luck may be understood as a type of error. If the predictive coding framework is correct, then minimising errors is the only task of the mind. In what regards luck, perhaps the very task of the mind is to minimise the effect of luck on thought and action. Maybe it can still happen, but we should not expect it to happen to a degree that threatens our conception of morally responsible action. More importantly, holding people responsible does not need to be independent of chance.

Rather, chance is what responsible agents are faced with, in their attempts to act in the world. So, minimising the effect of chance goes to the core of what responsible agency means.

The paper began with a rather bleak outlook. The existence of resultant luck suggests that people are not responsible for their actions. However, there is reason to think that ordinary action approximates an ongoing, many-shots process of constant error minimisation. And this seems to be consistent with a plausible cognitive framework whose very purpose is to maximise control by minimising errors, luck, and uncertainty.

References

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bermúdez, J. L. (2000). Personal and sub-personal; A difference without a distinction. *Philosophical Explorations*, *3*(1), 63-82.

Bermúdez, J. L. (2020). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, *128*(9).

Björnsson, G., & Persson, K. (2012). The explanatory component of moral responsibility. *Noûs*, *46*(2), 326-354.

Björnsson, G. (2017). Explaining away epistemic skepticism about culpability. *Oxford studies in agency and responsibility*, *4*, 141-162.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Clark, A. 1989. *Microcognition: Philosophy, cognitive science, and parallel distributed processing*, London: MIT Press.

Clark, A. (2011). “What scientific concept would improve everybody’s cognitive toolkit?” In: *Edge*. url: https://www.edge.org/response-detail/10404.

Clark, A. (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science”. In: *Behavioral and Brain Sciences* 36, pp. 181–253.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Colombo, M. (2013). Constitutive relevance and the personal/subpersonal distinction. *Philosophical Psychology*, *26*(4), 547-570.

Colombo, M. (2017). Social Motivation in Computational Neuroscience, in *The Routledge Handbook of Philosophy of the Social Mind*, (ed.) J. Kiverstein, New York: Routledge: 320–40.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, *60*(23), 685-700.

Davidson, D. (1974). Psychology as philosophy. In *Philosophy of psychology* (pp. 41-52). Palgrave Macmillan, London.

Dennett, D. C. (1969). *Content and Consciousness*. London: Routledge and Kegan Paul.

Dennett, D. C. (1984). Control and Self-Control. In *Elbow Room: The Varieties of Free Will Worth Wanting* (pp. 55–80). The MIT Press.

Dretske, F. I. (1967). Can events move?. *Mind*, *76*(304), 479-492.

Frankfurt, H. (1993). What we are morally responsible for. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility*. Ithaca, NY: Cornell University Press.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417-458.

Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, *13*(7), 293-301.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.

Friston, K. (2018). Does predictive coding have a future?. *Nature neuroscience*, *21*(8), 1019-1021.

Galton, A., & Mizoguchi, R. (2009). The water falls but the waterfall does not fall: New perspectives on objects, processes and events. *Applied Ontology*, *4*(2), 71-107.

Hartman, R. J. (2017). *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness*. Routledge.

Hartman, R. J. (2020). Indirectly free actions, libertarianism, and resultant moral luck. *Erkenntnis*, *85*(6), 1417-1436.

Hohwy, J. (2013). *The predictive mind*. New York: Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259-285.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209-223.

Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese*, *199*(1), 29-53.

Holton, R. (2016). Thinking ahead: Review of “Surfing uncertainty” by Andy Clark. *The Times Literary Supplement*.

Hornsby, J. (2000). Personal and sub-personal; A defence of Dennett's early distinction. *Philosophical Explorations*, *3*(1), 6-24.

Hoyer, P. O., & Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in neural information processing systems* (pp. 293-300).

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580-593.

Khoury, A. C. (2012). Responsibility, tracing, and consequences. *Canadian Journal of Philosophy*, *42*(3-4), 187-207.

Khoury, A. C. (2013). Synchronic and diachronic responsibility. *Philosophical studies*, *165*(3), 735-752.

Khoury, A. C. (2018). The objects of moral responsibility. *Philosophical Studies, 175*(6), 1357-1381.

Lang, G. (2021). *Strokes of Luck: A Study in Moral and Political Philosophy*. Oxford University Press.

Latus, A. 2000, Moral and Epistemic Luck, *Journal of Philosophical Research*, 25, 149–172.

Lewis, D. (1988). Desire as belief. *Mind, 97*(387), 323-332.

Lewis, D. (1996). Desire as belief II. *Mind, 105*(418), 303-313.

Ludwig, K. (2010). Adverbs of Action and Logical Form. *A Companion to the Philosophy of Action*, 40-49.

Marušić, B., & Schwenkler, J. (2018). Intending is believing: A defense of strong cognitivism. *Analytic philosophy*, *59*(3).

Nelkin, D. (2013). Moral luck. *The Stanford Encyclopedia of Philosophy.* In E. Zalta (ed.) https://plato.stanford.edu/archives/win2013/entries/moral-luck/.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87.

Ridge, M. (1998). Humean intentions. *American Philosophical Quarterly*, *35*(2), 157-178.

Sanborn, A. N., & Chater, N. (2017). The sampling brain. *Trends in cognitive sciences*, *21*(7), 492-493.

Steward, H. (2012). Actions as processes. *Philosophical Perspectives*, *26*, 373-388.

Thompson, M. (2004). Apprehending human form. *Royal Institute of Philosophy Supplements*, *54*, 47-74.

Thompson, M. (2008). Naïve action theory. *Life and Action*, *2010*, 85-148.

Thomson, J. J. (1989). Morality and bad luck. *Metaphilosophy*, *20*(3/4), 203-221.

Wolfson, B. (2012). Agential knowledge, action and process. *Theoria*, *78*(4), 326-357.

Zimmerman, M. J. (1987). Luck and moral responsibility. *Ethics*, *97*(2), 374-386.

Zimmerman, M. J. (2002). Taking luck seriously. *The Journal of Philosophy*, *99*(11), 553-576.

Concluding Remarks

We all face the world with an unavoidable degree of uncertainty, independently of whether the universe is deterministic or indeterministic. Control of thought and action is subject to this uncertainty. We each have a limited perspective on the world, and even on our own mentality. However, there is a kind of freedom we have within this restricted domain. Contra serious objections, this kind of freedom may be sufficient to realise the sort of control that is necessary to be morally responsible.

I began this dissertation by turning the focus on the agent. Specifically, the first article argued that for an event to be considered an action, it is vital that the agent has contributed to it in a meaningful way. This notion of agentive contribution is an indicator of the degree of control one has, which in turn determines whether one is morally responsible. I then argued for a psychologically plausible way to understand the exercise of responsibility-level control, against the challenges presented by a functional/mechanistic explanatory framework.

In the second article, I presented a version of the problem of luck which affects agents regardless of the truth of determinism. Mental luck is founded on the idea that our epistemic perspective of the world is inevitably limited, such that free decisions may be indistinguishable from decisions that are a matter of luck. While that may not be objectively true, it is subjectively so from the perspective of an ordinary agent. Thus, it is impossible to tell which of our decisions are prone to mental luck. The sombre conclusion of the article suggests that mental luck poses a serious challenge to the realisability of the control condition and the possibility of free action.

Finally, the third article began with a different luck-related worry, this time resultant luck. Cases like *drunk driving* and *assassin* apparently demonstrate that people are -at most- only morally responsible for their mental willings, but not for their actions and their consequences. However, such cases are founded on the assumption that actions are events. Once we drop that assumption and conceive of actions as ongoing goal-directed processes involving many-shots, cases of resultant luck are by no means representative of ordinary action. To support my claim, I aligned this alternative ontology with a naturalistic cognitive architecture, according to which the mind is primarily involved in minimising errors through Bayesian updating or active inference. The *many shots* view of action significantly diminishes the force of prominent cases of resultant luck.

The third article provides support for the claims made in article 1, while it also serves as a counterargument to the worry about mental luck from article 2. If actions are open processes that involve many-shots, it is possible to think of control for action as an ongoing process of conforming the world to our desires. Accordingly, if error minimisation is the primary task of the mind, we should consider the possibility that luck is another type of error. Our cognitive capacities may well be attuned to exercise control in a way that minimises the effects of luck, within this environment of subjective uncertainty and limited information. I do not provide such a sweeping argument in the article. My hope is that this will form part of future work on this topic. That said, the foundations of freedom in uncertainty are present here, spelled out in terms of a plausible account of responsibility-level control.

Uncertainty, luck, chance, and limited information have traditionally been viewed as threats to our control and freedom. This thesis challenges this predominant perception by proposing a credible account of freedom in uncertainty. Once we have accepted that we live in an environment of subjective uncertainty, the objective nature of the universe has little real-world significance for our actions. We will never know if our next endeavour will be successful if our latest decision was truly free or the result of some chance process. Even if we did, we would surely harbour some inexplicable and uneasy hesitation about the future. Yet, uncertainty does not mean restraint. Throughout this work, I have attempted to explore the foundations of the sort of freedom we can reasonably expect to have, even within this limited domain. That freedom may be restricted in many ways, but we have reason to think our mind has ways to exercise a degree of control over its environment and enable us to act freely, for the most part. Of course, that involves a comprehensive investigation which far exceeds the boundaries of my dissertation. I believe such work will contribute to a better understanding of human cognition and action in the literature and encourage more contact between philosophy and the cognitive sciences. I have every intention to engage in this work in the future.