Responsibility and Ambivalence

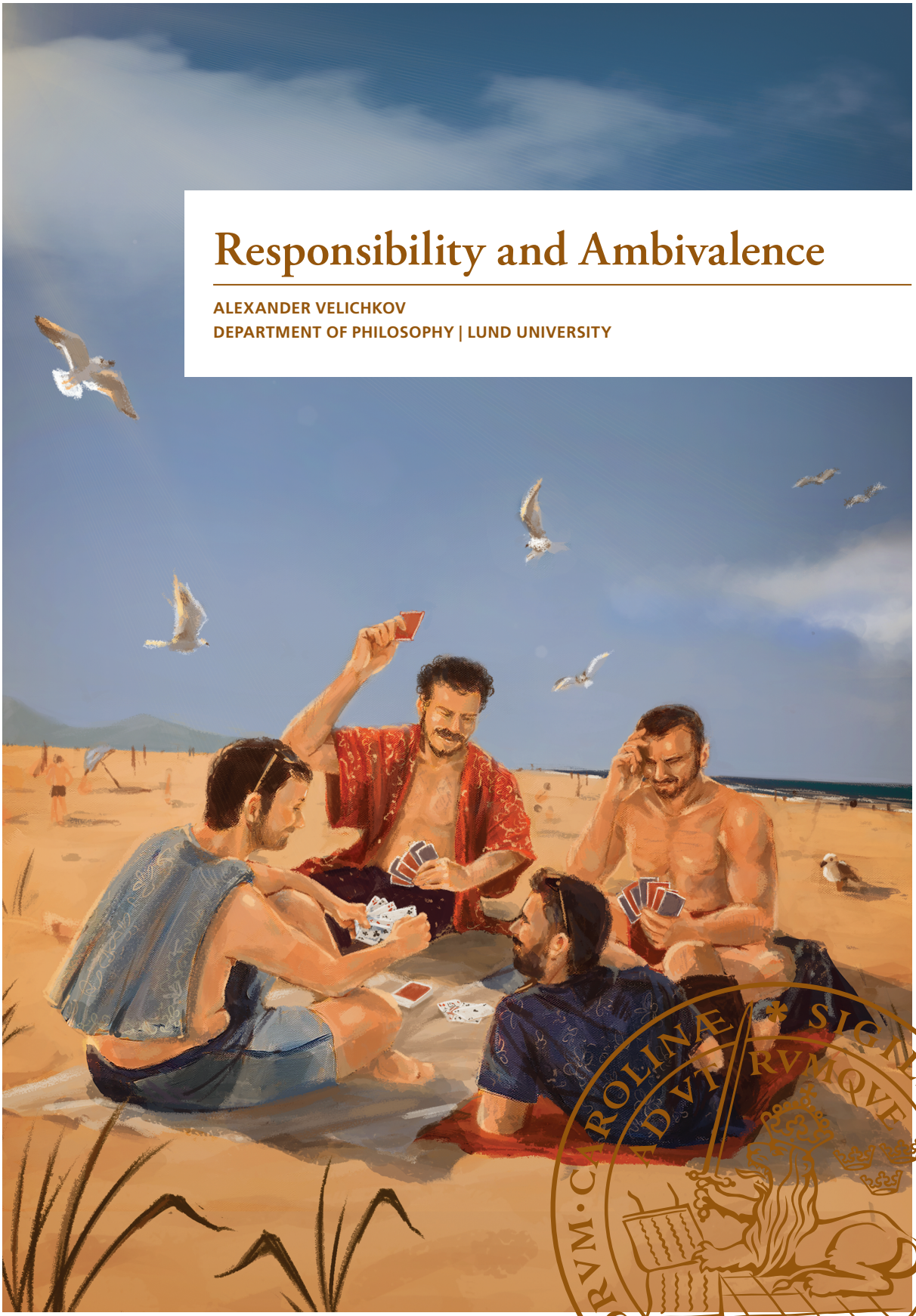
Velichkov, Alexander

2023

Document Version:
Publisher's PDF, also known as Version of record

Responsibility and Ambivalence

ALEXANDER VELICHKOV
DEPARTMENT OF PHILOSOPHY | LUND UNIVERSITY



Responsibility and Ambivalence

Responsibility and Ambivalence

Alexander Velichkov

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Joint Faculties of Humanities and Theology at Lund University to be publicly defended on the 14th of October 2023 at 11.00 in LUX:C126, Helgonavägen 3, Lund.

Faculty opponent: Professor David Shoemaker

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation **Date of issue:** 14th of October, 2023

Author(s): Alexander Velichkov

Title and subtitle: Responsibility and Ambivalence

Abstract:

I use the concept of *ambivalence*—the state of being faced with a choice that cannot be resolved without sacrificing something of value—to approach five contemporary debates in the philosophy of moral responsibility: (1) psychopathy, (2) free will, (3) the emotion of guilt, (4) regret and indirect moral luck, and (5) moral demandingness. Rather than arguing for one theory or another, acknowledging ambivalence paves the way for resolving these debates by reconciling the opposing sides.

Key words: responsibility, free will, ambivalence, guilt, regret, psychopathy

Language: English

ISBN (print): 978-91-89415-92-8
ISBN (digital): 978-91-89415-93-5

Number of pages: 144

Responsibility and Ambivalence

Alexander Velichkov

LUND
UNIVERSITY

Coverphoto: *Come on, Kasparov!* by Stephanie Stamenova:

Faculty: Humanities and Theology

Department: Philosophy

ISBN (print): 978-91-89415-92-8
ISBN (digital): 978-91-89415-93-5

Printed in Sweden by Media-Tryck, Lund University

Lund 2023

To Sven and Stephie

Table of Contents

Acknowledgments

I would like to express my heartfelt gratitude to my supervisors Björn Petersson, András Szigeti, and Olle Blomberg. Not only did they provide me with invaluable and resolute support with my PhD studies, but they also helped in making these four years the happiest in my life, and the ones I have grown the most in. I could not have asked for better mentors.

I would also like to thank my office mates, my academic co-pilgrims. Anton Emilsson, thank you for being my brother-in-PhD—we began this journey together on the 1st of September 2019 and have been companions through the difficult and the easy times since then. Marta Johansson Werkmäster and Jakob Werkmäster, thank you for Music Fridays; Gloria Mähringer, thank you for being always so supportive; Mattias Gunnemyr, thank you for being an epitome of kindness; Marianna Leventi, thank you for your great and truly unique sense of humour; Robert Pál-Wallin, *brate*, thank you for being my indispensable emotions-support, both philosophically and otherwise; Jiwon Kim, thank you for being a source of inspiration; Martin Sjöberg, thank you for the warm-heartedness that I encountered at the start of every workday. May LUX:B532 remain in history as the cradle of the Lund Circle of Philosophy.

I owe much to Paul Russell, who gave me guidance and detailed comments on my work. Paul was also the one who suggested the title *Responsibility and Ambivalence*. Thank you, Agnés S. C. Baehni, for helping me see beauty in philosophy and in life; I owe a lot of Chapter 4 to you. I would like to thank Frits Gåvertsson for being a philosophical inspiration and for guiding me in the world of ancient ethics. I also wish to express my deep gratitude to Sophie Grace Chappell, who acted as opponent at my final seminar. Her insightful comments and encouragement were indispensable for improving my thesis.

My monograph has benefitted immensely from comments by the participants at the Higher Seminar in Practical Philosophy at Lund University and the seminars held by the Lund Gothenburg Responsibility Project (LGRP). In addition to my aforementioned colleagues, I wish to thank David Alm, Henrik Andersson, Monika Betzler, Gunnar Björnsson, Eric Brandstedt, Ellen Davidsson, Dan Egonsson, Hadi Fazeli, Andrés Garcia, Frits Gåvertsson, Anders Herlitz, Ingvar Johansson, Yuliya

Kanygina, Jenny Magnusson, Shervin MirzaeiGhazi, August Olsen, Erik Persson, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, Paul Russell, Signe Savén, Jakob Stenseke, Matthew Talbert, Daniel Telech, and Patrick Todd. Each one of you has shaped my thesis, and I was lucky enough to become great friends with some of you.

I greatly appreciate the comments on my work that I have received by Julien Deonna, Fabrice Teroni, and the *Thumos* research group; Stefaan Cuypers and Benjamin De Mesel at KU Leuven; Theron Pummer and the participants of CEPPA 2020 at the University of St. Andrews; and the participants at ECAP10.

My philosophical work was made possible by the administrative work of David Alm, Anna Cagnan Enhörning, Tobias Hansson Wahlberg, Petter Johansson, Martin Jönsson, Anna Östberg, Tomas Persson, and Annah Smedberg-Eivers. Thank you so much for your help and support and for being patient with me.

My travels and education during my PhD have been generously funded by Stiftelsen Erik och Gurli Hultengrens fond för filosofi, Fil. Dr. Uno Otterstedts fond för främjandet av vetenskaplig forskning och undervisning, Dagmar Perssons fond, and the Lund Gothenburg Responsibility Project, for which I am deeply grateful.

I had the honour of serving as chair of the *Lund Philosophical Society* between 2021 and 2023. I wish to thank Niklas Dahl, Ellen Davidsson, Anton Emilsson, Ulf Gallbo, Frits Gåvertsson, Mattias Gunnemyr, Jiwon Kim, Marianna Leventi, Jenny Magnusson, Gloria Mähringer, Aaron Milz, Max Minden Ribeiro, Joel Parthemore, and Robert Pál-Wallin for their devout work to the Society. It was a pleasure working together.

I greatly enjoyed my time in Sweden thanks also to my dear friends and colleagues at the department outside of practical philosophy: Thibault Boehly, Mark Bowker, Niklas Dahl, Max Minden Ribeiro, Fredrik Österblom, Sami Stedtler, Trond Tjø*s*theim, Matt Tompkins, Melina Tsapos, Gabriel Vogel, and Anton Wrisberg.

The PhD is a challenging time, and I fared well in large part thanks to the friendship of Nia Mihaela Aleva, Roberta Dimana Aleva, Kamen Brestnichki, Konnie Brestnichka, Alexander Georgiev, Gorie, Roel van Herk, Viktor Hortman, Svetlin Ivanov, Alisha Leeming, Marijn Mado, Rossen Manev, Steffany Mihaylova, Angel Novoselski, Nikolina Novoselska, Kriso Petrov, Nikolai Radkov, Lukas Reimann, Daniela Stamenova, Stephan Stamenov, Kristin Stoyadinova, Anka Supej, the publicity department of Turkish Airways, Toon de Vries, Julika Wolf, and Filip Zahariev. Special thanks to Martin Angelov, Anton Dessov, Evgeny Penchev, and Alexander Radev for unintentionally featuring in the painting on the front cover.

I wish to thank my parents Zdravka Velichkova and Rumen Velichkov for teaching me resilience and for giving me the gift of education. My thanks also to my brother

Vladislav Velichkov and his family Yana Vassileva, Yavor Velichkov, and Adelina Velichkova, for being such a source of love. I hope you all know that I would not have gone this far without you.

Last, but not least, I would like to thank Stephanie Stamenova for being the sun and joy in my days, for the wonderful painting *Come on, Kasparov!* that is on the front cover, and for the photo on the back cover. Thanks also Sven for being my role model of equanimity.

Metaphysics is still needed by some, but so is that impetuous *demand for certainty* that today discharges itself in scientific-positivistic form among great masses—the demand that one *wants* by all means something to be firm (while owing to the fervour of this demand one treats the demonstration of this certainty more lightly and negligently): this is still the demand for foothold, support—in short, the *instinct of weakness* that, to be sure, does not create sundry religions, forms of metaphysics, and convictions but does—preserve them.

—Friedrich Nietzsche, *The Gay Science*

Chapter 1. Introduction

1.1 Moral responsibility as a set of existential problems

1.1.1 The structure of existential problems

Human emotions have evolved to deal with the problems specific to human existence. Many of us intuitively fear high places, spiders, snakes, sharp teeth, and claws; we are disgusted by dirt and rot; we are sad when we lose close ones or objects of value. A highly social species like us also needs social emotions to keep up group cohesion and cooperation. We feel sympathy for the vulnerable, righteous anger at offenders, guilt and shame about our own misdeeds, admiration for virtuous people, pride at our own achievements, and, as I shall go into detail below, intuitively care about social equity. All these emotions motivate behavior that in general serves to meet the challenges of being human.

Besides our emotions, we also have the capacity to reason, which allows us to extend emotional evaluation to abstract objects or ones that are not present in our immediate environment. We may fear not only the lightning in the sky but also the void of death. Insecurity may arise out of doubting the existence of the external world. Some of us may become bored by the whole drab familiar universe rather than just by this eventless afternoon. The pointlessness of the task at hand may frustrate us but so can the pointlessness of our lives as a whole.

The emotions, expanded by our rational engagement with the world, thus shape the human condition by generating what can be called problems of human existence. These are problems *of* human existence rather than *about* human existence because they are not mere intellectual puzzles; they pose frustrations that detract from our well-being. They are of *human* existence because they concern aspects of existence that we humans generally share, such as our mortality, and the pointlessness, insecurity, and injustice permeating our lives.

It is an interesting open question how much we can suppress the natural activity of these two capacities—reason and the emotions—that together produce existential

problems. Human life is difficult to imagine without these capacities, as they are crucial for dealing with the concrete, particular challenges that we face every day. To stop fearing death, we need to stop fearing or stop thinking about death, and neither seems like an easy thing to do.

Besides suppressing our emotions or our reasoning, another strategy for releasing ourselves from the burden of existential problems is to interpret the universe in a way that dissolves them. Comforting stories that attempt this have cropped up in cultures all over the world. Some are associated with religion and mythology: death is not scary because there is an afterlife, divine beings ensure that everyone will ultimately get their just deserts, and human existence serves some greater purpose. "Serious" scientific inquiries, in contrast, are supposed to disclose the true extent of our existential predicament by not sugarcoating reality. Scientific thought, however, is by no means free from the optimistic assumptions of some religious worldviews (Nietzsche, 2003, p. 110). Philosophy, and especially moral philosophy, is no exception. Moral theories suffer from a tendency to present ethical life as tidier, fairer, and easier than is evidenced by ordinary experience. Bernard Williams—the philosopher who has perhaps done most in the 20th century to critique this comforting side of moral philosophy—argues that the tragedies of ancient Greece are closer to the uncomfortable truths of ethics than many of the great philosophical works (Williams, 1993, pp. 163-164).

1.1.2 Ambivalence and incommensurability

Some existential problems, like those in the examples above, come about because a desire or need cannot be satisfied in the world as it is. However, some existential problems have a different structure: they arise because two or more of our desires or needs are inconsistent with each other, in the sense that both cannot be simultaneously satisfied. A curious example is the clash between one's aversion to death and one's aversion to boredom; perhaps one can either have immortality or a relatively boredom-free life, but not both (Williams, 1973b).

An especially troubling subset of these problems of inconsistent desires or needs is encountered in the moral sphere. Sometimes, two moral values (like equality and liberty) cannot be simultaneously realized, and so one of the two must tragically be sacrificed. On a very simplistic picture, resources can be redistributed to achieve social equality, but this will infringe on people's property rights; or people can keep their property, but this will lead to an unequal society. Something valuable will be lost either way.

We can use "ambivalence" as a technical term to denote the state of being faced with a choice that cannot be resolved without such a (moral or non-moral)[1] value remainder (Coates, 2023, Chapter 2). When a person is ambivalent, her choice is difficult not because, for instance, she does not have enough information or is too akratic to act. Rather, her choice is hard because she knows that something of value will be lost whatever she does.

Choices that cannot be resolved without leaving a value remainder can be divided into two broad kinds. The first kind is such that one option is better than the others, despite its being costly. All things being equal, an agent must prefer it. For instance, it is better for a field surgeon to save a person's life than to alleviate the harmless but intense pain of someone else, assuming that she cannot do both.

The second kind is comprised of choices where one course of action is no better than another, and so practical deliberation cannot recommend either. Which one of two drowning people should I save, assuming that I cannot save both? These choices often leave us dumbfounded, as the values that can be realized may be of equal worth, or perhaps, if they cannot be reduced to a common measure, incommensurable. Philosophers have pointed out other possibilities too, such as the values' being on a par or being clouded by vagueness (see e.g., Andersson & Herlitz, 2021, 2022; Broome, 1997; Chang, 2002; Rabinowicz, 2009). We can call these *hard cases* and the corresponding attitude—*hard ambivalence*. Whatever the appropriate decision procedure is in hard cases, it cannot, by definition, exclusively rely on weighing the relevant values against each other.

Much of moral philosophy takes ambivalence in the moral sphere as the starting point for reflection—as a problem that can and must be resolved—rather than its endpoint. Moral theories like deontology and consequentialism provide decision procedures that aim to resolve such ethical quandaries. The problem with many of these theories is that they do not resolve ambivalence as much as explain it away by discounting one (or a few) of the values that constitute the conflict by deeming it irrational or confused. For consequentialists only consequences matter, and so no ethical weight is given to the intrinsic moral worth of actions (like the intrinsic wrongness of torture). Conversely, deontologists take certain actions as impermissible even if failure to perform them would bring about horrible consequences. Each theory, then, gives a clear (and opposing) answer to the dilemma between torturing one person and letting five people die; and neither allows for the possibility of facing this choice with (hard) ambivalence.

[1] Throughout the monograph, I will rely on an intuitive distinction between moral and non-moral (e.g., aesthetic, athletic, epistemic) values.

The well-explored reductivism of moral theories such as deontology and consequentialism for the sake of providing a useful decision procedure (see e.g., S. G. Chappell, 2022, Chapter 2) is a symptom of the wider problematic tendency of moral philosophy to overreach itself. With the ambitions of a law-discovering natural scientist, the philosopher often tries to work out clear general solutions to ethical problems: psychopaths are not blameworthy for their actions, anger is always irrational, Artificial Intelligence systems should be treated like moral agents, etc. (As Alan Watts says, today's philosophers would wear a white lab coat to work if they could get away with it.) It may be underwhelming, especially to the general public, to prescribe ambivalence (especially hard ambivalence) for difficult ethical issues. Yet, often this may be the only honest verdict. Some might worry that moral philosophy would be rendered useless once it gives up on the search for decision procedures that would yield unambivalent results. This is not so. The methods and insight of philosophy are irreplaceable, at the very least to map out ethical conflicts by identifying the values involved and rooting out prejudices and errors. An accurate, truthful account of any given ambivalence must be the starting point for dealing with it.

The search for unambivalent solutions does more than conceal ethical truth and the true extent of existential problems in the sphere of the ethical; it tends to create *deadlocked debates*. Think again of deontology and consequentialism. When a problem arises because two valid considerations come into conflict (e.g., that consequences matter and that actions also matter intrinsically), and the assumption in the discussion is that only of the two must be the correct one, then it is only natural that two theories will appear, each supporting one of the two considerations and discounting the other. Since each theory is half-correct, neither side will convince the other and they will remain at an impasse. The concept of ambivalence therefore has the potential to *reconcile*. When a debate has this deadlocked structure, showing that each side correctly identifies a valid ethical consideration paves the way for synthesizing them.

1.1.3 Responsibility

A budding subfield in moral philosophy, the philosophy of responsibility has its share of deadlocked debates that can be usefully approached by exploring the possibility of ambivalence. Although there are insightful accounts of ambivalence in responsibility (e.g., Shoemaker, 2015; Watson, 2004a, 2004b), there is a need to pursue the phenomenon further. The aim of this monograph is to fill this gap by characterizing a number of ambivalences having to do with five central problems in the philosophy of responsibility: psychopathy (Chapter 2), free will (Chapter 3), the emotion of guilt (Chapter 4), regret and indirect moral luck (Chapter 5), and moral demandingness (Chapter 6). As I mentioned above, moral ambivalence is often taken by philosophers

as a starting point and then "resolved" by not giving due weight to the normative forces that give rise to it. My approach will be to go the other way around: I will investigate the normative forces at play in responsibility and show how, if acknowledged, they lead to ambivalence.

What are these normative forces? Chapters 2, 3, and 4 are primarily devoted to exploring certain ambivalences about how to hold others responsible. These arise from the conflict between, on the one hand, *the sense of fairness* and, on the other, the various intuitions about *responsibility norms* governing how we should respond, emotionally and through actions, to the ethical worth of an agent's behavior. According to the sense of fairness, people should not be treated differently (measured in burdens and benefits) in virtue of features that are outside of their control. It is unfair, for example, to receive higher pay because of one's skin or eye color. Fairness is a powerful consideration that permeates many aspects of ethical life that go beyond responsibility. Distinct from it are the norms of our responsibility practices, such as the ones guiding the fittingness of the responsibility emotions, or the so-called "reactive attitudes" (P. F. Strawson, 2003), like resentment, indignation, gratitude, disappointment, guilt, and agential regret. Besides the fittingness of the emotions themselves, responsibility norms also govern these emotions' natural expression through actions like overt blame and praise, apology, forgiveness, reward, and punishment. Sometimes, an agent may be appropriately resented, and even blamed and punished, for what is to some extent outside of his control, and this is when the sense of fairness and responsibility norms collide, resulting in ambivalence. For instance, to give a glimpse of what is to come: I will argue that it is unfair, and at the same time appropriate, to resent, blame, and punish psychopaths.

Chapters 5 and 6 will use the concept of ambivalence in a somewhat different way. Rather than characterizing ambivalences about holding others responsible, I will explore the *agent's experience of being responsible for making choices that merit ambivalence*. Put differently, the focus will shift from hard choices about holding responsible to being responsible for making a hard choice. In Chapter 5, I will take Williams's (1981b) Gauguin case as a starting point to explore the idea that whether one views one's past morally wrong decision as all-things-considered justified depends on how one happens to transform as a person in the course of one's life. Finally, in Chapter 6, I will show that often (in contemporary times, quite often), many are faced with a distressing ambivalence about the choice between leading a personal life and the demanding moral project of helping others in need.

In each chapter, I will suggest a way for reconciling two opposing (sets of) theories into a synthesis that acknowledges ambivalence. Since these ambivalences depend on features that are, in general, characteristic of human life—our ultimate lack of control,

guilt, regret, and the overdemandingness of morality—they may be called existential problems. In all likelihood, many of us will encounter them in the course of our lives.

1.2 Theoretical foundations

Let me introduce the concepts that will play a central role in the rest of the text: moral luck, free will, moral responsibility, the fittingness of emotions, and fairness. Each one of these concepts is complex and controversial enough to merit a monograph on its own. I will only discuss them briefly in this section, as a preliminary for delving into the philosophy of responsibility, but I will continue to elaborate on them and engage with some of the debates surrounding them throughout the monograph.

1.2.1 Moral luck

A tenet of moral responsibility is the strong intuition that one is responsible only for what is within one's control. Yet, numerous examples from everyday experience seem to contradict this control condition on responsibility. Determining the extent to which factors beyond the agent's control influence her blame- or praiseworthiness is known as the problem of moral luck.

Thomas Nagel (1979) set the philosophical agenda for this major ethical problem with a useful taxonomy of the different kinds of moral luck:

> *Resultant/outcome luck* is luck in the outcome of one's actions. An unlucky drunk driver who hits a crossing pedestrian appears more blameworthy than an equally drunk driver who luckily encounters no pedestrians on her way home.
>
> *Circumstantial luck* is luck in one's circumstances, which affects whether one's agency gets expressed in a blame- or praiseworthy way. The Nazi responsible for horrible acts would have been morally innocent if he had happened to settle in Argentina before the 1930s.
>
> *Constitutive luck* is luck in one's agential structure (made up of such things as character traits and capacities). A person's aggressive, selfish, or cowardly behavior may be ultimately traced back to her genes and upbringing.
>
> *Causal luck* is luck in the causes of one's actions; in other words, this is what is standardly thought to be the problem of free will. In a deterministic universe, all of our actions are products of causal chains that flow outside of our control, and so our agency appears to be ultimately subject to luck.

This list is not exhaustive. New kinds of moral luck continue to be proposed. As I will argue in Chapter 5, Williams's (1981b) famous Gauguin case is not, as is standardly thought, an instance of simple resultant luck (Nelkin, 2021), but rather of luck in how one transforms throughout one's life (what can be called "transformative luck" (Lang, 2018)). Another candidate is *associative luck* (Telech, 2022): luck in responsibility in virtue of one's attachments to other persons, as when one's grandfather is a Nazi officer or when one is the citizen of a morally flawed state.

Furthermore, it is not even clear that all of the items on Nagel's list are distinctive: causal luck appears to collapse into a combination of constitutive and circumstantial luck (Latus, 2001).[2] The problem of causal luck, i.e., free will, just is the problem that we are ultimately lucky in the inward causes of our actions (how we are constituted as agents) and the outward causes of our actions (the situations we find ourselves in). Whether it is true that causal luck just is constitutive plus circumstantial luck depends on how deep constitutive luck goes. Some think of it as luck in the set of character traits that one is given and must then suppress or endorse through the power of one's will, as if one is dealt cards with which one decides how to play the game of life (e.g., Enoch & Marmor, 2007, p. 430). According to others, however, this is an unduly restricted account of constitutive luck because one's will, or the very capacities to suppress or endorse one's character traits, is subject to luck too (Levy, 2011). Luck not only deals the cards, *but also the decision procedure for playing them*. On this deeper view, the problem of free will indeed seems to collapse into the problems of constitutive and circumstantial luck.

When Williams introduced the phrase "moral luck", he expected it to "suggest an oxymoron" (Williams, 1995c, p. 241) because, on a dominant conception, moral responsibility is supposed to be one realm of life free from luck. My arguments in this monograph will show that the tension contained in the oxymoron is not illusory. Moral luck denotes an area of ethical life marked by ambivalence.

1.2.2 Free will

Perhaps the oldest problem still in use by philosophers, the problem of free will, though centered around human agency, actually encompasses a myriad of questions about God, destiny, the physical laws of the universe, quantum mechanics, mind-body dualism, the phenomenology of action, moral and legal responsibility, and the emotions. The literature is colossal, as there are numerous debates in the vicinity of free

[2] I offer a possible explanation of why Nagel thought of causal luck as a distinct kind of moral luck in Chapter 3.

will, and it is not always clear if philosophers are on the same page about what the problem under discussion even is. In this subsection, I will offer the conceptual mapping of the central debate that I find most useful for my purposes in this monograph.

To anchor the discussion, free will can be defined in the following way:

> *Free will* is the unique ability of persons to exercise the strongest sense of control over their actions necessary for moral responsibility. (McKenna & Pereboom, 2016, p. 6)

Unfortunately, defining free will in terms of moral responsibility only takes us so far because the latter is a multifaceted and ambiguous concept. Still, some excellent philosophical work in the last few decades has helped specify the various meanings of "moral responsibility" in more and more detail. Moral responsibility seems to have a lot to do with the fittingness of the reactive attitudes (McKenna, 2012; Shoemaker, 2015; P. F. Strawson, 2003; R. J. Wallace, 1994), but for some, emotional responses are not all there is to it. Another much-discussed contemporary proposal is that the kind of moral responsibility relevant for free will should be spelled out as deserving good/bad treatment solely in virtue of performing morally right/wrong acts with sensitivity of their moral status (Pereboom, 2021, p. 11).

Having more or less homed in on the relevant capacity of free will, what is it that threatens its existence? Traditionally, the main threat was considered to be *determinism*, the thesis that every event is necessitated by antecedent events together with the laws of nature. With regard to determinism, there are two possibilities:

> *Compatibilism:* free will is compatible with determinism.
>
> *Incompatibilism:* free will is incompatible with determinism.

It is usually implicitly assumed that compatibilism affirms the existence of free will, though it does not strictly speaking entail it. Incompatibilism in turn can be subdivided into two other theories:

> *Hard determinism:* free will is incompatible with determinism, determinism is true, and so there is no free will.
>
> *Libertarianism:* free will is incompatible with determinism, but determinism is false, and there is free will.

Notice that libertarians do not just reject determinism. They also affirm some kind of contra-causal (i.e., not causally determined) free will. It would not matter much for us

if the only indeterministic processes occurred, say, in some far-off cosmic gasses. The relevant indeterminism must be constitutive of our agency. There are several competing accounts of how this can happen: our actions are not caused (e.g., Lowe, 2008; McCann, 1998; Palmer, 2020), they are caused by mental events (e.g., Clarke, 2003; Ekstrom, 1999; Kane, 1996), or they are caused by agents (e.g., Chisholm, 1976; Griffith, 2010; O'Connor, 2000).

However, the debate has shifted in response to the idea that free will is endangered not just by determinism. Non-determinism is also incompatible with free will because it implies that our actions, if not wholly determined by antecedent events and the laws of nature, result instead from pure chance (like the random firings of neurons). Therefore, regardless of whether determinism or indeterminism is true, we do not have ultimate control over our actions (Pereboom, 2014, 2021). Moreover, libertarian theories are not only false, but their concept of contra-causal free will, however it is spelled out, is incoherent (G. Strawson, 1994). This family of views is known as:

> *Hard indeterminism*: free will does not exist regardless of whether determinism or indeterminism is true.

Unless stated otherwise, I will use "hard indeterminism" and "free will skepticism" interchangeably.

Note that *free will skepticism* should be distinguished from *moral responsibility skepticism.* The latter is broader than the former, as it refers to all views that reject the idea that human beings can be morally responsible, regardless of whether they reach that conclusion via worries about free will. New kinds of moral responsibility skepticism keep being proposed, such as skepticism based on our epistemic limitations in knowing the true causes of our actions (Rosen, 2004).

In relation to the shift to hard incompatibilism, compatibilism now often unofficially implies its rejection, and implicitly means something like "Free will is compatible with a naturalistic account of the world". Compatibilists reject free will skepticism by pointing out capacities that make up free will and that can be reasonably assumed to exist, such as the ability to endorse or disavow one's desires (Frankfurt, 1971; Watson, 1975) or to respond to reasons (Fischer & Ravizza, 1998; Wolf, 1987, 1994). To act freely, for many compatibilists, just is to act without coercion, constraint, or compulsion. One is unfree when one is, for instance, in chains, hypnotized, under the influence of drugs, or suffering from some severe mental disorder.

At this point in the debate, free will skepticism and compatibilism are at an impasse. Both are supported by strong intuitions, and one of the big current challenges is to see

how much, if at all, they can be reconciled. I will address this issue in Chapter 3 by building on an approach to free will (and morality) put forward by Bernard Williams (1981b, 1993, 1995a, 2011) and further developed by Paul Russell (2017, 2022).

1.2.3 Moral responsibility

Key to finding one's way in debates about moral luck (including the free will problem) is clarifying the concept of moral responsibility. The study of the nature of moral responsibility in Anglophone academic philosophy is prolific, though the discipline is relatively young, beginning in the 1960s as an offshoot of the free will debate, largely due to P. F. Strawson's (2003) influential "Freedom and Resentment" (originally published in 1962). Now it encompasses many more issues having to do with the nature and norms of praise and blame.

As I mentioned above, the central worry with moral luck is that it seems inappropriate to hold agents responsible for what is outside of their control. In this thesis, I will pursue the thought that this control condition on responsibility is in part supported by the intuition that it is *unfair* to hold agents responsible for what is outside of their control. The extent of this unfairness depends on how much the agent is *harmed* or *benefitted* by being held responsible, which brings up the question: what are responsibility responses and how harmful or beneficial are they?

Philosophers have identified a big cluster of responsibility responses that interact with each other in complex ways. Here, I will mention the ones that will play a prominent part in the monograph. A core response is *aretaic appraisal*. On Watson's (2004b) influential view, this "face" of responsibility is called attributability and is essentially an evaluation of a person's character. It is difficult to imagine skepticism about such aretaic appraisals, as people obviously act in ways that are cowardly, kind, wicked, generous, etc. For most people, it can hardly be called inappropriate to evaluate a person who knowingly and intentionally tortures animals for fun as cruel. Aretaic appraisals are accompanied by reactive attitudes like admiration and disapproval.

The other face of responsibility, for Watson, is *accountability.* Certain actions for which the agent is accountable merit not only aretaic appraisals but also a variety of other responses. These include reactive attitudes that involve dispositions to treat others in unwelcome ways, such as resentment and indignation (Watson, 2004b, pp. 278-279), as well as overt responses like meting out punishments and rewards, and demanding an apology and an acknowledgment of the wrong committed (see e.g., Bennett, 2002; McKenna, 2021). Unlike aretaic appraisals, many of these overt responses are clearly burdensome or beneficial for their recipient, and so the question of their fairness is more pertinent.

It is no surprise that theorists who defend the idea that an agent can be held responsible for behavior that is outside of her control tend to locate the essence of responsibility in relatively mild responses, like private and overt aretaic appraisals and reactive attitudes (e.g., Adams, 1985; Hieronymi, 2004, 2008; Smith, 2005, 2015; Talbert, 2012). In contrast, authors who emphasize the unfairness of moral luck focus on punishment and reward. As I mentioned in the previous subsection, free will skeptics in particular spell out moral responsibility in terms of the idea that a person deserves reward or punishment solely in virtue of his good or bad actions performed with sensitivity to their moral status. This idea is known in the free will debate as "basic desert" (Pereboom, 2014, 2021), but variants of it have been explored under the name *retributivism*. I will come back to basic desert and retributivism in Chapters 2 and 3. My investigation of ambivalence in responsibility is in part made possible by not limiting the discussion to only certain responsibility responses. A complete picture of responsibility requires looking at the beneficial and the unbeneficial, as well as the harmless and the harmful.

1.2.4 The fittingness of emotions

I suggested above that the structure of emotions sheds light on the structure of existential problems, and it is no different for the problems of moral responsibility, many of which revolve around the norms guiding the reactive attitudes.

The philosophical study of emotions has accumulated a vast literature. I cannot do full justice to it here; instead, I will note a few key features relevant for discussing their normativity. A useful model for understanding emotions is to see them as comprised of several parts. (Frijda, 1986; Jacobson & D'Arms, 2006; Shoemaker, 2015, p. 21). An emotion's *appraisal* is an evaluation and interpretation of its object. Fearing a dog involves appraising the dog as dangerous. An emotion's *affect* is how the emotion feels to the person experiencing it. One feels one's heart racing and short of breath when in a state of fear. *Emotional motivation* can be usefully separated into two parts. An emotion's *action tendencies* are the specific behaviors that it prepares one for. Action tendencies of fear include freezing and fleeing the dangerous object. An emotion's *goal* is the state that satisfies the emotion. The goal of fear is threat avoidance (D'Arms & Jacobson, 2023, p. 105).

When it comes to the normativity of emotions, there are multiple dimensions on which an emotion can be correct vis-à-vis its object. Intuitively, there is a kind of fit between an emotion's appraisal and some specific kinds of objects. Fear is correct when directed at a dangerous attacking dog. One technical term for this kind of relation is *fittingness,* and it is supposed to be distinct from moral or prudential reasons to experience an

emotion (D'Arms & Jacobson, 2000; Rabinowicz & Rønnow-Rasmussen, 2004). Thus, it might be morally wrong and imprudent to feel jealous of a more successful colleague, and yet it might be a fitting emotional response. Moreover, a fitting emotion is not necessarily the epistemically warranted emotion (Echeverri, 2019). My fear of a highly realistic toy cobra is unfitting even if no one can criticize me for believing it is the real thing. Keeping epistemic, prudential, moral, and fittingness dimensions of normativity separate is crucial for investigating cases of ethical ambivalence, as they often arise out of a clash between different kinds of norms.

While intuitive on the surface, there are many aspects of emotional fittingness that remain controversial. Are fittingness norms relative to an individual, a particular culture, or hold universally? Should fittingness be spelled out in terms of correct representation of the object of an emotion, as when fear represents the dog as dangerous (Milona, 2016; Roberts, 2003; Tappolet, 2016)? Or should it be understood instead as correct engagement with the object, as when fear makes one avoid the dangerous dog, have an increased heartrate around it, etc. (D'Arms & Jacobson, 2023; Deonna & Teroni, 2012, 2015, 2022; Mitchell, 2021; Mulligan, 2007; Müller, 2017)? How distinct are fittingness norms from prudential or moral ones? Thankfully, the lack of agreement on these and other issues does not render the concept of fittingness unusable; it will do.

1.2.5 Fairness

Fairness is, unfortunately, at least as elusive as the previous concepts. By fairness, I mean equitable treatment within some system of distributing burdens and benefits. There are various ways to specify what equity consists in, and none is unanimously accepted by philosophers. I will rely on the following principle:

> *FAIRNESS*: people should not be treated differently, in terms of burdens and benefits, for what is outside of their control.

This principle is one possible construal of a sense of fairness, which, evidence suggests, is innate to human beings; it is culturally universal, spontaneous, and found in young children (Baumard, 2016; Debove, Baumard, & André, 2017). Interestingly, an intuitive aversion to social inequity (a primitive form of the sense of fairness) is observed in species for whom cooperation is key to survival, such as chimpanzees (Proctor, Williamson, de Waal, & Brosnan, 2013), capuchin monkeys (Brosnan & de Waal, 2003), and dogs (Range, Horn, Viranyi, & Huber, 2009). The correlation between some sense of social equity and cooperation gives a hint about the evolutionary origin and function of the human sense of fairness.

Adhering to the principle of fairness is a key element of impartial treatment. It requires that if two people both do the same amount of work, they should be paid equally, and factors such as sex or skin color that are outside of their control are irrelevant. Fairness is the backbone of much of ethical and political life, including distributive and retributive justice, and even of some relations within the family. Yet, it does not guide all human interactions (think of romantic love), nor is it an overriding moral force; sometimes, an unfair outcome might be all things considered preferable.

Ironically, on a popular interpretation, David Hume (2009)—one of the founders of sentimentalism, the idea that morality is ultimately grounded in the emotions—believed that fairness is a useful social convention (i.e., an "artificial virtue") rather than an innate moral concern.[3] His sentimentalist contemporary Francis Hutcheson (2008; see also Mounce, 1999, p. 97), in contrast, argued the opposite. One may get quite different ethical results depending on whether one follows Hume or Hutcheson on this. If Hume is right, then fairness can easily be abandoned when alternative social conventions would prove more conducive to cooperation. However, if Hutcheson is right, then there is some ethical cost to unfair outcomes because they are intrinsically morally wrong. Hutcheson is closer to the truth if we are to trust our intuitions and the available evidence, and I will assume that he is right: fairness is one among the many genuinely moral concerns that shape ethical life.

1.3 Overview of chapters

Chapter 2. Psychopathy

The responsibility of psychopaths (i.e., morally incapacitated agents) is a primary example of a problem meriting ambivalence. On the one hand, psychopaths perform atrocious acts. On the other, their moral incapacities prevent them from appreciating their morally wrong actions as morally wrong. Among philosophers, there are those who hold that psychopaths can be morally responsible and those who think they ought to be exempt. I will argue that both sides have a point. A psychopath is indeed neither like an ordinary wrongdoer nor like a cognitively incapacitated agent who is out of touch with reality. Psychopaths' morally wrong actions may merit certain responsibility responses; yet, it is unfair to harm them through these responses, since their wrong actions are due to incapacities that are ultimately beyond their control. An ambivalent attitude—it is unfair and yet appropriate to respond negatively to psychopathic wrongdoing—is the truthful response to the tragic fact that a person can be evil through no fault of her own.

[3] For a more elaborate interpretation of Hume's views on fairness, see Queloz (2021b, Chapter 4).

Chapter 3. Free will

In this chapter, I will assume that free will skepticism is in part correct: psychopathy is just a salient example of the more general fact that all our moral flaws and merits are ultimately caused by forces outside of our control, such as our genes and upbringing. However, I will also assume that certain compatibilist accounts are also partly correct in that human beings are capable of voluntary action, which grounds many of our responsibility practices. In this chapter, therefore, I will not *argue* for a position in the free will debate. Instead, I will present a reconciliatory alternative on which free will skepticism and compatibilism are compatible. Each theory correctly identifies a set of moral concerns that conflict with each other and entail an ambivalence analogous to the one in Chapter 2: it is ultimately unfair to harm or benefit all agents in virtue of their responsibility (because of ultimate luck), though there are retributivist and instrumental reasons (grounded in voluntary action) to do so.

Chapter 4. Guilt

Chapters 4 and 5 turn away from problems of ultimate control and look instead respectively at the self-directed responsibility emotions of guilt and agential regret. In Chapter 3, I argue against the Standard Account of guilt, which states that guilt is fitting and/or deserved only if the agent is blameworthy, and I offer a novel account of guilt, the Moral Debt Account, which can make sense of fitting guilt without blameworthiness. An agent may have a moral debt, fittingly marked by guilt, that disrupts a relationship even though she is not blameworthy. Ambivalence here arises if we suppose that such guilt is *deserved*. If it is, then others would be licensed to impose it on the agent, and this brings up the issue of fairness again. It would be unfair, and yet deserved, to experience guilt for innocent behavior.

Chapter 5. Regret and indirect moral luck

At this point in the text, the reader might be tired of reading about fairness, and so in Chapter 5 I will approach moral luck from a different direction. I will interpret Williams's (1981b) puzzling claim in his seminal "Moral Luck" that rational justification (and somehow, because of this, moral justification) is subject to luck. The crux of Williams's argument revolves around personal transformation: the process of changing one's value system during one's life. I will offer an interpretation that makes sense of the idea that an agent's moral responsibility is not affected by how he transforms as a person, and yet this transformation affects whether his actions would be viewed by him as preferable from a broader, all-things-considered perspective. From a later standpoint, an agent like Williams's Gauguin might view his choice with ambivalence, as meriting moral blame and yet as overall preferable.

Chapter 6. Moral demandingness

In Chapter 6, I will approach the question "How much ought one sacrifice for the sake of morality?" as a problem of responsibility. I will challenge the widespread assumption that there is a neat line dividing the morally required from the morally supererogatory (i.e., the good but not required). I will argue that many agents living in an overdemanding moral landscape, like the one of today, face a difficult ethical conflict between their own autonomy and helping those in need. Sometimes, it is permissible to choose one's personal life over a moral cause and vice versa, but both choices may come with an ethical cost: one must either sacrifice one's autonomy and personal projects or must otherwise fail along a moral dimension, taking on the burden of guilt and disappointment for choosing oneself over others. There is no easy way out of this ambivalence.

Chapter 2. Psychopathy

2.1 Introduction

There is strong empirical evidence that psychopathy is a developmental disability that leads to immoral behaviour, such as lying, manipulating, and harming others (for an overview, see Blair, Hwang, White, & Meffert, 2016). Certain brain areas in psychopathic individuals appear to be impaired, and this correlates with a diminished ability to experience, among other things, empathy and guilt. Psychopathy is therefore narrower than what the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition* (DSM-5; American Psychiatric Association, 2013), terms Antisocial Personality Disorder, which may subsume a diverse range of individuals who exhibit antisocial tendencies. A psychopath is not just any particularly callous and irresponsible person; she is someone who has a developmental moral incapacity.

The nature of this moral incapacity is still being investigated, and so some qualifications are in order. Because psychopaths tend to fail at distinguishing moral from conventional norms (Blair, 1995; Blair & Cipolotti, 2000), I will assume that they are unable to *recognize* certain moral reasons as (*de re*) moral reasons rather than able to recognize them but unable to be motivated by them. I doubt that much hangs on this for the present study, as in either case the relevant end result is that psychopaths suffer from an inborn incapacity to act morally for the right kind of reason. Moreover, psychopathy might not affect one's whole moral compass. There is evidence that psychopaths perform badly in endorsing care-based moral norms, such as "Do not cause pain to others", but not when it comes to disgust-based ones, such as "Do not eat human flesh" (Aharoni, Antonenko, & Kiehl, 2011; Glenn, Iyer, Graham, Koleva, & Haidt, 2009). Finally, developmental psychopathy might be a matter of degree. By psychopath, I will mean someone who is on the far end of the spectrum and is unable to appreciate moral reasons as moral reasons. My conclusions will have to be altered to some extent to cover less severe cases.

Psychopathy is a good entryway into exploring ambivalence in responsibility, as it vividly illustrates the clash between fairness and responsibility norms, which, as I will argue, permeates much of our responsibility practices.

Some form of ambivalence (not necessarily in the technical sense in which I am using the term (see §1.1.2)) seems to be a natural response to psychopathy. Already in the 19th century Philippe Pinel, who was the first physician to study people with psychopathic traits, described his patients as suffering from "insanity without delirium" to capture their distinctively moral incapacities as opposed to deficiencies in mind or intellect (Kavka, 1949). In more recent times, Neil Levy (2007), echoing Pinel, calls the psychopath a "sane madman" (p. 129), and Dana Nelkin (2015) observes that "when we think about psychopaths' actions, and the pleasure they sometimes take in inflicting harm, it seems we are correct in blaming them; when we think about their deficits, it seems we are incorrect to do so" (p. 358).

Theories of psychopaths' responsibility explain this tension in different ways, and they range from one extreme to the other. There are views that psychopaths cannot be blameworthy at all (Levy, 2007, 2014; Nelkin, 2015; R. J. Wallace, 1994; Wolf, 1987), that they can be blameworthy, but not in the central "accountability" sense (Shoemaker, 2015; Watson, 2011), and that they can be fully blameworthy (Scanlon, 1998; Talbert, 2008, 2012, 2014). These disagreements appear to be at least in part due to the fact that responsibility responses that usually go together—such as negative reactive attitudes, aretaic evaluations, the making of demands, punishment, and protesting the wrongness of the action—come apart in the case of psychopaths; they cannot get "the whole package" of blame. Some of these responses become inappropriate, and a lot of the philosophical work on the topic is on establishing which, if any, can be retained. Depending on what one takes to be central or essential to blame, one gets different results. If what is essential is the making of demands for moral acknowledgment and apology, for example, then blame is inapt when directed at psychopaths (Watson, 2011); but if blame is taken to be about expressing negative reactive attitudes and protesting wrongful actions, then psychopaths are as blameworthy as ordinary wrongdoers (Talbert, 2008, 2012, 2014).

The purpose of this chapter is to bring together and build on the insights of existing philosophical work in order to provide a novel account of the ambivalent responses to psychopathic wrongdoing. I will argue that philosophical accounts of psychopathy have so far failed to focus on an ethically

salient, perhaps the most ethically salient source of ambivalence: the conflict between negative reactive attitudes and other negative responses on the one hand and considerations of fairness on the other. I will also suggest one possible reason for why this has been overlooked. Compatibilist authors, who have produced a big portion of the recent work on psychopathic agency, are theoretically constrained from allowing worries about an agent's lack of ultimate control over her actions to enter their theories. If one admits that a reason for mitigating blaming responses in the case of psychopaths is that their incapacities arise out of forces beyond their control, one can easily generalize this into free will skepticism.

I will first explore the appropriateness of ascribing virtues and vices, and of directing reactive attitudes to psychopaths (§2.2), before turning to the appropriateness of making demands on them (§2.3). I will then argue that the psychopathic agent merits ambivalence, as considerations of fairness come into conflict with retributive and instrumental reasons to respond negatively to his bad acts (§2.4). Finally, I will conclude that many prominent theories of responsibility are wrong to bundle together cognitive incapacities and normative incapacities because the two affect responsibility in different ways (§2.5).

2.2 Reactive attitudes and aretaic evaluations

Because of their moral incapacities, psychopaths are more likely to cheat, manipulate, torture, rape, murder, exploit, and abuse, or in other words, to commit some of the worst possible acts of which humans are capable. There are many difficulties involved with measuring psychopathy, but according to some estimates psychopaths make up only around 1% of the general population (Sanz-García, Gesteira, Sanz, & García-Vera, 2021), while they comprise as much as 25% of prison inmate population in the US (S. J. Morse, 2013, p. 41). Many people certainly call psychopathic wrongdoers cruel and callous and feel negative reactive attitudes like resentment and indignation towards them. Some philosophers, however, question whether these aretaic evaluations and reactive attitudes are appropriate.

In this section, I will review and argue against attempts to show that psychopaths do not merit these responses. This is not yet a discussion about whether one should have or express reactive attitudes to psychopaths *all things considered*. How the blaming emotions conflict with other ethical and prudential considerations will be the topic of the following sections. Here, the more modest

task is to make it plausible that psychopaths can exhibit traits that are at least *pro tanto* resentment- and indignation-worthy.

Some authors argue that psychopaths' incapacities prevent them from *manifesting ill will*, which is standardly taken to be what makes the blaming emotions fitting. Consider this analogy offered by Neil Levy:[1]

> Suppose there is a kind of harm that is objectively morally relevant, but of which we are ignorant. Suppose, for instance, that plants can be harmed, and that this harm is a moral reason against killing or treading on them. In that case, many of us are (causally) responsible for a great many moral harms. But it is false that we express contempt, ill-will, or even *moral* indifference to these plants. (Levy, 2007, p. 135, emphasis in original)

If the analogy between unknowingly harming plants and psychopathic wrongdoing holds, then we would indeed be wrong to respond with negative reactive attitudes to the latter. Blaming a psychopath would be as inappropriate as blaming someone who burns large amounts of fossil fuels but is non-culpably ignorant of the effects of her actions on the environment.

However, is psychopathy analogous to ignorantly and harmfully treading on plants? Not quite. Matthew Talbert (2014) notes that the two cases differ in a crucial respect (pp. 283-284). The plant-killers are not aware that they are *causing harm*. In contrast, psychopaths know that they are causing harm. They know that inflicting pain harms people, and in fact sometimes use this to their advantage, by inflicting pain or threatening to do so in order to reach their goals. Rather, what psychopaths cannot grasp is that an action's causing harm is a moral reason for refraining from that action. They are thus a lot closer to non-psychopathic wrongdoers than the plant-treading example makes them appear, and this seems to bear on our intuitions. To take the above example again, imagine that rather than being ignorant of the effects of her actions on the environment, the fossil-fuel burner does know this but just does not care.

In what sense, however, do psychopaths believe they are harming others? Closely connected to the plant-treading argument is what Levy (2014) calls the argument from content. There is evidence that psychopaths have deficits in mental time travel (MTT), the ability to project oneself into the future or the

[1] A similar case is also found in Shoemaker (2011, p. 625).

past. MTT is what makes long-term planning and the pursuit of projects possible for human beings. Levy maintains that being an *autonomous person* is largely dependent on MTT because "[a] person's capacity for pursuing plans and projects is the capacity for autonomous action: for action that imposes a shape on that person's life" (Levy, 2014, p. 362). If psychopaths are deficient in MTT, they cannot see themselves or others as fully autonomous beings, or in other words, *as persons*. They cannot intend harming a person *qua person* and therefore their actions do not have the *moral content* necessary for blameworthiness (Levy, 2014, p. 352). A psychopath cannot understand that by, say, killing someone, he is also ending the life of an autonomous person who has projects and plans:

> One psychopath candidly confesses that he felt no worse about the people he harmed than about squashing a bug (Hare, 1999, p. 33). From his perspective, there is little difference: neither bug nor human is understood as an autonomous being. (Levy, 2014, p. 363)

In response to Levy's argument from content, one may question whether the evidence supports his claim that psychopaths fail at long-term planning in such a way that they cannot understand autonomy and personhood. It should be noted that psychopaths are rational agents who are capable of at least some forms of planning and deliberation. They effectively execute short-term plans, and they have projects, goals, and dreams, even though they might be less likely—or even quite unlikely—to pursue them with commitment and consistency. As Levy himself notes, failure in long-term planning is not a necessary feature of psychopathy; there are many high-functioning psychopaths who are as successful as other members of society (Levy, 2014, p. 363). Many psychopaths have problems, on the face of it, not because they are unable to *conceive of* long-term plans, but rather because they *act impulsively*: "Impulsivity and a failure to plan ahead is a diagnostic criterion for antisocial personality disorder, the DSM (IV-R) category that is closest to psychopathy" (Levy, 2014, p. 361). It is doubtful that systematic impulsivity is evidence that one cannot understand what long-term plans and projects are. A plausible alternative explanation is that psychopaths can understand and make long-term plans, but *simply do not care about executing them*. That is, a response parallel to the plant-treading case could be made in response to Levy's argument from content: it is one thing to not know what projects and plans are, and another to know this but not care about them. Failure in MTT could be spelled out in the following way: psychopaths' deficiencies in empathy prevent them from empathizing with their own future

selves, which in turn prevents them from *executing* long-term plans. In other words, they treat themselves and their projects like they treat others—with disregard (Watson, 2013). More evidence is therefore needed to establish that psychopaths lack an understanding of long-term planning (and thereby autonomy) that goes beyond mere impulsivity and lack of interest or care.

Yet, even if we grant that psychopaths do not grasp the full extent of the harm they cause to a person by ruining her long-term projects, one could make the case that this would still only exempt them from some of the moral wrongs that they commit. This is because there are other moral values that seem to be independent from that aspect of personhood. It seems plausible, for example, that inflicting bodily harm, torture, emotional abuse, stealing, and cheating are wrong in themselves, regardless of whether this affects the victim's long-term projects. For instance, the primary reason against killing a toddler does not appear to be that one is preventing that toddler from one day growing into a person who has personal projects. In sum, even if Levy is correct that some psychopaths lack the imaginative capabilities to grasp long-term projects, and that furthermore they do so in a way that undermines their understanding of autonomy, his argument would only suffice to exempt them from certain acts. Besides violations of autonomy, psychopaths would still be blameworthy for the many other kinds of wrongs they commit.

Another challenge against the claim that negative reactive attitudes are fittingly directed at psychopaths is Nelkin's (2015) argument from asymmetry. Nelkin maintains that psychopaths do not merit moral aretaic ascriptions and the emotions that go with them—one cannot properly say that a psychopath is, for example, cruel or kind. To make her case, Nelkin asks us to

> [i]magine a creature who lacks all moral understanding and does not see others' interests as fundamentally reason-giving, but who, along with beer and cigarettes, enjoys watching car chases and other people enjoying themselves. He doesn't care about other people in the sense of taking their interests to be reasons for acting. He just enjoys seeing them having a good time. In such a case, in my view, we would be reluctant to describe the case as one in which the creature is kind. Similarly, it is not clear why we should describe the psychopath as one who is cruel. (Nelkin, 2015, p. 363)

Furthermore, this pro-social psychopath

> might enjoy being the agent of another's pleasure more than simply watching it. These things don't add up to kindness unless one does something for the sake of someone. I am not exactly sure what the parallel of doing something for the sake of someone is in the case of cruelty or acts manifesting contempt. But the mere taking of pleasure in the pain, and even in being the instrument of pain, does not entail cruelty or contempt. (Nelkin, 2015, pp. 367-368)

Intuitively, the pro-social psychopath that Nelkin describes is not kind because of his lack of moral understanding. And unless we can explain why there is an asymmetry in the requirements for moral understanding between moral virtues and vices, we must conclude that psychopaths cannot be cruel either.

Talbert (2021) responds to Nelkin by granting that there must be a symmetry between virtues and vices when it comes to moral understanding. Instead, he challenges the premise that the pro-social psychopath does not merit positive reactive attitudes:

> . . .the pro-social psychopath might come close to giving us what we minimally desire from strangers. We don't have to call this "kindness," but it is a kind of amiableness, and I suspect that something like gratitude would not be out of place as a response to it. (Talbert, 2021, p. 1240)

I doubt that the pro-social psychopath merits positive reactive attitudes like gratitude. If I knew that someone, say, gave me presents just because he found it enjoyable to be the agent of my pleasure, I would not call him kind. Even though I might be happy to see him because of the presents he gave me, I would not feel grateful *to him*; and I would not feel like I should praise him for his taking selfish pleasure in gift-giving. After all, there is nothing about me as a person that he cares about. When he gets tired of taking pleasure in this way, he would easily move on to another hobby. If I call him amiable, it seems like this would be a shallow compliment not accompanied by genuine gratitude.

It could be granted to Nelkin, therefore, that the pro-social psychopath is not kind and does not merit positive reactive attitudes. Rather, one should deny that asymmetry is a problem. In fact, the asymmetry is only to be expected. This is because the distinction Nelkin is pointing out *just is* what distinguishes certain moral vices from the virtues. The virtue of kindness involves responding to the

moral reasons surrounding caring for others. Conversely, being cruel, callous, and inconsiderate just is to fail taking such moral reasons into consideration.

Nelkin says that she is "not exactly sure what the parallel of doing something for the sake of someone is in the case of cruelty or acts manifesting contempt", and there is good reason for her uncertainty. There is no parallel (or at least not for many moral vices). Nelkin seems to suppose that a cruel act is one in which an agent first recognizes another's interests as non-instrumentally morally valuable and then disregards them. But in many everyday acts of cruelty, what appears to happen is that the cruel agent simply does not care about the other's interests in the first place. To be cruel and inconsiderate just means to omit taking a human as a morally valuable being, and to treat her at best as a means to some further goal, as when the psychopath extracts pleasure by hurting others. It is precisely the psychopath's capacity to *rationally* target someone as a means to some further end (like her own pleasure) that makes psychopathic agency chillingly cruel and more offensive than that of predators like wolves.[2]

Notice that Nelkin's argument reveals an overlooked aspect of psychopaths' responsibility: they might be incapable of genuine acts of kindness. Their incapacities preclude them from being fitting targets of moral praise, though, if I am correct, their wrongful acts evoke fitting resentment and indignation. Psychopaths are thus in a tragic position: they can be morally vicious, but not morally virtuous.

So far, I have presented and noted the weaknesses of some of the most compelling arguments against having negative reactive attitudes towards psychopathic wrongdoing. I now turn to two considerations that speak in favor.

First, as Talbert suggests, it would be very difficult to distinguish the reasons for which a psychopath acts from those of a non-psychopathic wrongdoer:

> . . . suppose that we have two wrongdoers (*A* and *B*) who both cheat an old woman out of money and who both truthfully report that they did

[2] One might be concerned that if psychopaths merit the blaming emotions, so should animals—after all, both psychopaths and animals lack moral competence—but then something must have gone wrong in the argument. As Talbert (2014) and Watson (2011) point out, however, psychopaths, in contrast to animals, are rational agents and language users. They can cheat, manipulate, lie, exploit, enjoy other people's suffering, and use people as means to their ends. Furthermore, in virtue of their rationality, psychopaths can in principle fulfill social roles: they can be family members, colleagues, and citizens.

> so because they wanted to see "the look of panic on an old lady's face." Next, suppose we find out that wrongdoer *A*, but not *B*, could have formed a morally preferable judgment about the normative status of the consequences of her action. *A* could have judged—though she did not—that the prospect of the old lady's suffering was a sufficient reason to not cheat her. . . How does the fact that *A* had psychological access to a morally preferable, but entirely counterfactual, instance of moral awareness make the quality of her action, or the quality of her will, more malicious, or more morally significant than *B*'s? (Talbert, 2014, p. 285)

In many cases of cruel or inconsiderate behavior, it appears that the agent does not first appreciate the moral status of her victim and then decide to overrule it. Rather, the victim's moral status does not enter the agent's deliberation in the first place (Talbert, 2021, p. 1242). She views people as objects, as numbers, as means to an end. Psychopathic wrongdoing thus seems indistinguishable from non-psychopathic wrongdoing as far as the actual sequence of events goes.

If the actual sequence of events is the same in two cases, where two agents perform the same acts *for the same reasons* (i.e. they exhibit the same quality of will), it is doubtful that emotions like resentment would be fitting in one case and not in the other. (Recall that what is at stake here is not whether these emotions are all things considered appropriate—I will get to this later on.)

Secondly, there is a line of reasoning that might convince theorists who are friendly to the idea of evolutionary explanations of emotions. The negative reactive attitudes like resentment and indignation evolved in human beings to fulfill certain social functions, the most obvious one of which is the regulation of harmful behavior (Aharoni & Fridlund, 2013; Mackie, 1982). It is plausible that getting angry at wrongdoers and responding with blaming, threatening, ostracizing, and punishing has been the primary way to suppress violence within a group. If this is true, then why would psychopathic wrongdoing not be the *paradigmatic case* of the resentment-worthy rather than an exception? Some of the strongest negative feelings of which humans are capable are responses to acts like the deliberate murder of children, and for good evolutionary reasons. These are acts that psychopaths, due to their incapacities, perform with an untroubled conscience. Imagine that we find out one day that some political leaders who have committed genocide to further their political goals were psychopaths. Would the anger felt by many at their crimes against humanity turn out to have been misplaced? At least as far as evolutionary reasoning goes, *pace* Levy, to think

that a human being can be squashed like a bug is the essence of ill will that merits resentment rather than a consideration that excuses.

Admittedly, the arguments in this section might not sway everyone's intuitions. I have, however, opened up the space for the idea that one may fittingly respond with blaming emotions to psychopathic wrongdoing. My strategy will now be to explain that other major attempts to argue that psychopaths are not blameworthy fail even though they are based on sound intuitions. Much resistance to holding psychopaths responsible comes from the correct observation that there are reasons for stifling the overt expression of the negative reactive attitudes when directed at psychopaths. I now turn to the first kind of such inappropriate outward responses: the making of demands.

2.3 Demands and protest

Authors like Gary Watson (2011) believe that the psychopath does not merit the negative reactive attitudes because it is inapt to make certain kinds of demands of her, such as the demands for apology or moral acknowledgment. In Watson's influential terminology, this means that psychopaths are not accountability-responsible, though because one can ascribe virtues and vices to them, they are attributability-responsible (Watson, 2004b; see §1.2.3).

However, just because some *expressions* of the blaming emotions, or some acts associated with them, are inappropriate (like the making of demands) does not necessarily make the emotions themselves unfitting (D'Arms & Jacobson, 2000; Rabinowicz & Rønnow-Rasmussen, 2004). This is especially so if there are other expressions of these emotions that are appropriate. It is therefore best to keep the questions of blaming emotions and acts of blaming separate.[3] To take a standard example: I might have excellent reasons to conceal my fear from a vicious dog as it might feel it and become even more confident in its attack. Yet, fear is nevertheless the fitting response, since it tracks the dangerousness of the dog. In a similar way, making an overt demand of a psychopath might be senseless, but the blaming emotions themselves are not for this reason necessarily made unfitting; again, there might be other expressions of these attitudes besides demands that are completely sensible.

[3] For an elaboration of this critique of Watson, see Talbert (2014, pp. 289-ff).

If we put emotions to the side, Watson is nevertheless correct to say that some blaming practices are senseless when it comes to psychopathic agency:

> To be clear, my concerns about holding psychopaths accountable are not in the first place about "fairness." They are about the conceptual aptness of making a "demand" of a creature that is incapable of recognizing one's standing to make demands. Is this to say that holding the psychopath responsible is inapt in just the way that addressing a fencepost, for instance, would be? In some ways "yes," in some ways "no."
>
> Psychopaths are neither merely things nor brutes, however brutishly they might at times behave. They are influenceable to various degrees by threats, negotiation, and persuasion. Nevertheless, because they are incapable of the reciprocity that demanding and owing justification presumes, moral criticism is not only futile but senseless. Nothing they could do could be intelligibly construed as an apology or acknowledgment. (Watson, 2011, p. 314)

In this passage, "moral criticism" refers to the core practice of blaming as accountability. It is a kind of dialogue where one demands from the perpetrator that she gives an apology or acknowledges the wrong that she has done, and, Watson argues, in the case of the psychopath these demands are senseless.

Watson echoes a point made by Bernard Williams (1995b). In his discussion of internal reasons, Williams argues that a form of blame is inappropriate in a range of cases, which he calls *hard cases*. The agents in these cases cannot, from their subjective motivational set S (comprised of their desires, projects, and commitments), via a sound deliberative route reach the conclusion that they should perform an action ϕ. For instance, when we tell someone that they have a reason to be nicer to their wife, they may answer: "I don't care. Don't you understand? I really do not care" (Williams, 1995b, p. 39). There is nothing in their motivational set S that can make them find a (moral) reason to be nicer; they are beyond the blamer's rational reach.

Williams (1995b), however, is explicit that the hard case is only beyond *one form* of blame, what he calls "focused blame" (p. 40): by blaming an agent for a particular action in this manner, we point out to them and make vivid a reason they had in their motivational set S to perform ϕ. Alternatively, focused blame may function as a "proleptic mechanism" through which we *give* agents a reason to perform ϕ. Since people have a general interest in being respected by others,

they might acknowledge our blame and through this enter our community of shared moral values. Williams suggests that proleptic blame can be thought of as a form of retrospective advice. If we could have advised the agent not to ϕ before the action, we could then blame her for not ϕ-ing after the action. Psychopaths are clearly a subset of hard cases, and since it is obvious that one cannot advise a psychopath that they have a moral reason not to commit a wrong, they would be beyond this kind of focused blame too.

Yet not all forms of blame are inapt in the case of psychopathic agency. Just as Watson states that psychopaths can be called vicious, so Williams (1995b) says that the man who has no internal reason to be nicer to his wife may be called ”ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal, and many other disadvantageous things” (p. 39). There is also what Williams calls ”moralism”: the kind of blame ”that hopes by mere force to focus on the agent’s reasons a judgement that represents in fact only a rejection (perhaps an entirely justified rejection) of what he has done” (Williams, 1995b, p. 44). However, a justified rejection of the wrongful act need not have as its goal reaching (“by mere force”) the wrongdoer. It may also be conceived as a kind of *protest* by the victim. Through such rejection, a victim affirms her moral values, which has moral significance regardless of whether it has a chance to reach the conscience of the perpetrator (Talbert, 2012).

Furthermore, although the philosophical point is that it would be inapt to demand acknowledgment from the psychopath, in real life we might never be in a position to be sure whether an agent is really beyond our reach (Williams, 1995b, p. 43). In the passage quoted above, Watson makes a distinction between a demand’s being futile and its being senseless. It might be futile, for instance, to make demands of incorrigible racists and Mafiosi, but, he argues, there is a categorical difference between them and psychopaths, since the latter have *no chance* to be brought back to a moral outlook. There is nothing for them to return to (Watson, 2011, p. 319). Though this difference might exist in theory, in practice, we are epistemically limited, and moreover, as I mentioned above, psychopathy might be a matter of degree. We may, therefore, very often (or perhaps always) allow for the possibility of hope that the person being blamed would acknowledge the moral values she has violated. The distinction between futile and senseless demands might never be clearly drawn in reality. As Watson tells us, even the vicious murderer Robert Harris, who exhibited many

psychopathic characteristics, repented in the last years before his execution (Watson, 2004a, p. 259).

Let us take stock. I have argued that there are good reasons to believe that the negative reactive attitudes, as well as negative aretaic evaluations, can be fitting in the case of psychopathic wrongdoing. On the other hand, overtly blaming would be inappropriate if its goal is to make demands of apology or acknowledgment. Still, because of our epistemic limitations, we might very rarely, if ever, be in a position to ascertain with reasonable certainty that an agent is truly beyond the reach of a demand. Hope to receive acknowledgment might make blame apt even when directed at psychopaths. Finally, despite these obstacles to some forms of blame, others, such as blame in the form of moral protest, can nevertheless be legitimate.

2.4 Fairness and ambivalence

I have so far not said anything about how theorists explain the ambivalence (again, not necessarily in my technical sense of the term) one might feel towards psychopathic agency. For Nelkin, who maintains that psychopaths cannot be blameworthy at all (because they cannot properly be called e.g., cruel or inconsiderate), the ambivalence would presumably stem from a mismatch between systematic horrific acts and the non-responsible agents causing them. According to Watson (2011) and Shoemaker (2015), the ambivalence arises because psychopaths can be blameworthy in one respect, but not in another. Both authors agree that psychopaths are attributability-responsible—they can have moral vices—but their incapacities prevent them from being accountability-responsible, which means that they do not merit the central, focused blaming emotions of resentment and indignation. I have expressed doubts that these authors are successful in showing that psychopaths do not merit the focused blaming emotions. Still, I do agree with another aspect of Watson's analysis. The making of demands, which naturally accompanies the blaming emotions, might be inapt, and this might explain part of the ambivalence.[4]

All of these accounts of ambivalence, however, still overlook a powerful and ethically salient intuition that has to do with psychopathic agency. Since they

[4] The senselessness of one form of moral criticism is also pointed out as a source of ambivalence by Scanlon (1998, pp. 288-289).

are developmentally incapacitated, it appears *unfair* to sanction psychopaths through our responsibility practices. This does not mean that psychopaths do not commit wrongful acts or cannot be cruel, as the arguments by Levy and Nelkin attempt to show. Rather, the point is that they are set up to be cruel by their genes and environment. Through no fault of their own, they are incapable of recognizing moral reasons and thereby greatly disadvantaged in their ability to conform to moral rules.

If we acknowledge that the intuition of fairness is valid, what should we do about it? According to Levy (2007), the fact that psychopaths' bad behavior is ultimately beyond their control is sufficient reason to *fully exempt* them from moral responsibility. Vargas and Nichols put Levy's proposal in the form of a dilemma between accepting two competing and intuitively powerful theses:

> *UNFAIRNESS*: It is unfair to punish someone for a characteristic (or behavior) that was caused by factors outside the agent's scope of control
>
> *FAIRNESS*: It is (sometimes) fair to punish psychopaths (Vargas & Nichols, 2007, p. 153)

Which one should we choose? Then again, why do we have to choose one? Nichols and Vargas are presenting us with a false dilemma because there seem to be reasons both for and against punishing psychopaths. Their labeling is misleading, since these two theses are not both principles of fairness, or are only so in a very general sense. The two powerful intuitions at stake actually come from quite different sources.

UNFAIRNESS is indeed derived from an intuition of fairness, by which I mean the substantive principle that one should not be treated differently, in terms of burdens and benefits, for what is outside of one's control (see §1.2.5). People should not receive burdens in virtue of their developmental incapacities because these are outside of their control.

On the other hand, what Nichols and Vargas call *FAIRNESS* is an intuition that can come about in two ways. First, there is the retributive idea that wrongdoers, purely in virtue of their wrongful acts, are liable to punishment (or, more mildly, deserve to suffer the harmful effects of blaming, such as being ostracized or emotionally burdened (Bennett, 2002; McKenna, 2012, p. 141)). Retributive intuitions seem to be naturally connected to resentment (Shoemaker, 2015, p. 90; Watson, 2004b, pp. 278-279), and, as I showed in §2.2, psychopaths may

be the paradigmatic case of resentment-worthy agents. Retributivism is separate from the much broader substantive principle of fairness mentioned above, and one can consistently uphold the latter without the former. But even besides these retributive intuitions, one may still believe that, for instrumental reasons, psychopaths need to be punished, or at least confined. To prevent bad acts, society's laws must be upheld through a system of deterrence and confinement. We thus have at least three considerations at play:

> *UNFAIRNESS*: It is unfair to punish (or impose burdens on) someone for a characteristic (or behavior) that was caused by factors outside the agent's scope of control.
>
> *RETRIBUTIVISM*: it is right (or at least permissible) to punish or blame in a way that is harmful for a wrongdoer for their wrongful act.[5]
>
> *INSTRUMENTAL SANCTIONS*: Lawbreakers need to be punished or confined in order to sustain society.

Perhaps *INSTRUMENTAL SANCTIONS* does not necessarily need to come into conflict with *UNFAIRNESS*. Some might argue that a form of punishment or confinement is in the best interest of the psychopath, since that would help her integrate better in society. In the long run, such treatment is not a burden and is hence not unfair. This line of reasoning is doubtful and whether it is true in any particular case would ultimately depend on the extent to which the harms of punishment or confinement (e.g., ten years in prison or a mental health institution) outweigh the benefits gained (becoming a functioning member of society). Yet, even if *INSTRUMENTAL SANCTIONS* is in the end best for society as a whole but not for the psychopath, it is unlikely that we would judge that psychopaths can avoid some form of state-imposed unwelcome treatment. Though unfortunate, the threat of punishment might be the most effective way of deterring psychopaths from committing crimes.[6] The alternative, confinement in a mental health institution, would still be experienced as a burdensome loss of liberty for many.

---

[5] I will elaborate more on the kinds of retributivism in §3.2.2.

[6] Perhaps this expectation is too optimistic. As I discussed in §2.2, many psychopaths seem to be impulsive and bad at long-term planning, and so the threat of punishment might be an ineffective deterrent.

When we witness psychopathic wrongdoing, therefore, we may be torn. One may feel resentful and indignant at horrific acts performed rationally, systematically, and in cold blood—bring to mind a serial rapist and murderer like Ted Bundy—and one may want to blame, sanction, and punish in retribution. At the same time, on the assumption that the wrongdoer was born with an incapacity to see his actions as morally wrong, we may feel that it is unfair (in the sense of flouting a substantive principle of fairness) that he should suffer for this. Any one of us could have drawn the short straw and been in his place. Our negative reactive attitudes and the negative responses that they motivate clash with the concern we feel for our target when we realize that he would suffer unfairly because of his bad developmental luck. Moreover, even if one does not feel the pull of retributive intuitions, one may recognize that punishment or confinement of psychopaths is instrumentally necessary for sustaining society and for preserving security, which plausibly outweighs considerations of fairness. Responding negatively to a psychopath brings about retributive, or at least instrumental goods; abstaining from it preserves fairness. There is no way to have both. In essence, we should take the ambivalence that Nelkin points out at face value.

Given that the unfairness of responding negatively to psychopaths has already been pointed out in the literature, why has it been overlooked as a candidate for explaining ambivalence? Nichols and Vargas comment that

> . . .[*UNFAIRNESS*] is precisely the kind of principle that leads incompatibilists to conclude that no one is fully morally responsible (e.g., Pereboom 2001; Smilansky 2000). If incompatibilists are right about this, then psychopaths are by no means an exceptional case. The Unfairness principle, on this view, leads to sweeping exemptions for everyone. However, this would, we think, provide a powerful reason in favor of giving up on the [*UNFAIRNESS*] principle. (2007, pp. 154-155)

It is unsurprising that the debate over the responsibility of psychopaths has steered clear of intuitions of fairness. Most authors who address psychopathy are compatibilists (as Nichols and Vargas point out, psychopaths are not an exceptional case from a free will skeptic's perspective), and many of them might see themselves as theoretically constrained by their commitments in the free will debate. Standing by these commitments, however, comes at the cost of lacking the conceptual resources to draw the full picture of psychopathic agency.

Are compatibilist worries well founded? After all, what distinguishes genetic bad luck from other kinds of constitutive bad luck, such as luck in one's environment or in one's encounters with people and situations? As I will explore in Chapter 3, observations about psychopaths' responsibility do indeed generalize—however, crucially, they do not generalize into free will skepticism, or at least not into an unqualified kind. This is because it is not exemption that generalizes, but rather *ambivalence*. Non-psychopathic wrongdoers, just like psychopaths, may fittingly evoke *both* the blaming attitudes and intuitions *and* the sense that imposing the burdens of blame and punishment on them would be unfair due to their bad constitutive luck. A nuanced approach to free will that admits of insights from both compatibilism and skepticism is not barred by theoretical constraints from recognizing this profound ambivalence in responsibility.

To be fair to Nichols and Vargas, at the time of their writing, the debate surrounding free will skepticism had not yet resulted in a few important qualifications. In particular, only more recent work clarifies that what they call *UNFAIRNESS* targets only aspects of moral responsibility that constitute burdens or benefits, which leaves a variety of blaming responses appropriate (Pereboom, 2021, pp. 11-12). This should be noted in the case of psychopathy too. Blaming and punishment are unfair only to the extent that they are an overall burden to the psychopath. Thus, aretaic evaluations, having negative reactive attitudes, and some forms of moral criticism and protest are perhaps not in themselves unfair. Moreover, considerations of fairness need not be overriding. I leave it open how much instrumental and retributivist reasons ought to outweigh the substantive principle of fairness.

2.5 Normative and cognitive competence

In this section, I will pursue some important implications of my analysis of psychopathy for influential theories of responsibility according to which moral incapacities (like psychopathy) exempt the agent as much as certain cognitive incapacities (e.g., Fischer & Ravizza, 1998; Wallace, 1994; Wolf, 1987; 1990). My arguments in this chapter speak in favor of a more nuanced view: while certain cognitive incapacities do fully exempt the agent, moral incapacities like psychopathy lead to a complex responsibility response. The two kinds of conditions should not be bundled together. In this section, I will illustrate this difference by discussing Susan Wolf's (1987) popular Sane Deep-Self View.

According to Deep-Self theories of responsibility, an agent is responsible to the extent that she endorses her desires (Frankfurt, 1971; C. Taylor, 1976; Watson, 1975). Wolf (1987), however, argues that such views are incomplete because they lack an essential sanity condition. To see why, consider the case of JoJo:

> JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that tittle JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover' these are desires he wholly wants to have. When he steps back and asks "Do l really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (Wolf, 1987, p. 379)

Due to his upbringing, JoJo is incapable of responding to moral reasons, at least of a certain kind, and this makes him "insane". According to Wolf, since we have the intuition that he is not blameworthy—contrary to the verdict of Deep-Self views—the ability to endorse one's desires is insufficient for responsibility.

Wolf (1987) uses "sanity" in a special sense. It refers to the ability "cognitively and normatively to recognize and appreciate the world for what it is" (p. 383). Sanity, therefore, is a concept that puts together cognitive and normative competence. To defend her definition of sanity, Wolf appeals

> to the criteria for sanity that have historically been dominant in legal questions about responsibility. According to the M'Naughten Rule, a person is sane if (1) he knows what he is doing and (2) he knows that what he is doing is, as the case may be "right or wrong". (Wolf, 1987, p. 381)

For Wolf, the first condition in the M'Naughten Rule establishes whether the agent has the *cognitive* competence to form accurate *beliefs* about the world. The second condition, in contrast, establishes *normative* competence. It requires that the agent's *values* are appropriately guided by the world.

Wolf appeals to the wisdom of law, but her interpretation of the normative competence condition does not appear to be the dominant one in legal practice. Consider the clarifications of the M'Naughten Rule that a New York State judge is required to give to the jury when the defendant has raised an insanity plea:

> [W]ith respect to the term "wrong," a person lacks substantial capacity to know or appreciate that conduct is wrong if that person, as a result of mental disease or defect, lacked substantial capacity to know or appreciate either that the conduct was against the law or that it was against commonly held moral principles, or both. (New York State Unified Court System, n.d., p. 3)

This is already a broad construal of the M'Naughten Rule because it encompasses both moral and legal principles, and yet, the competence specified is very minimal. It only requires that the defendant was able at the time of the act to know that she was breaking a moral or legal rule. One only needs an ability to *form accurate beliefs* about what rules are and what the rules are to pass as sane. It does not require that the agent can *form correct values*. According to this interpretation, most psychopaths have sufficient normative competence to be legally responsible.

What would the M'Naughten Rule say about JoJo? Because of the nature of the example, JoJo is a very special case. Since he is a dictator, we can imagine that JoJo cannot go against the law in his country because he *is* the law. And perhaps we can also imagine that he is enveloped by subordinates who prevent him from knowing that according to most people in the world his actions are morally evil. Nevertheless, the elaboration of the M'Naughten Rule quoted above states that his normative ignorance must be "a result of mental disease or defect"; being surrounded by sycophants does not count. It is true that in JoJo's highly exceptional case, we might not be able to apply the M'Naughton Rule easily. The law was of course not designed to be applied to dictators who are above the law and isolated from the rest of humanity. Yet, the spirit of the M'Naughten Rule suggests that JoJo is sane enough to be legally responsible. Had he been a citizen rather than a dictator, he would have been legally culpable. JoJo's values may be faulty, but he is presumably not so severely cognitively incapacitated as to fail to understand rules.

Moving from legal to moral responsibility, there is also a difference between someone who is cognitively incapacitated to the extent that she fails by the

standards of the M'Naughten Rule and someone, like the psychopath, who knowingly and rationally commits crimes because of an inability to form the correct values. The former does not manifest ill will because she cannot even comprehend that people judge her acts as immoral. The latter, in contrast, knows that he is acting in ways that others take to be immoral but is simply not moved by these considerations. If my conclusion in the previous section is correct, he manifests ill will but brings up concerns of fairness. The psychopath, but not the cognitively insane, evokes ambivalence. These cases are importantly different, and the Sane Deep-Self View lacks the resources to distinguish between them. For Wolf (1994), a person is exempt to the extent that they fail to appreciate "the True and the Good". The theory thus puts together the severely cognitively incapacitated with agents like JoJo, "the slaveowners of the 1850s, the Nazis of the 1930s, and many male chauvinists of our fathers' generation" (Wolf, 1987, p. 382). Such a grouping might strike us as unintuitive, and the reason for this is that failing to see the True affects the complex web of responsibility responses differently from failing to see the Good.

Chapter 3. Free will

3.1 Introduction

In Chapter 2, I mentioned that psychopathy is just a salient example of the more general problem that human agency is ultimately subject to luck, i.e., the problem of free will. In this chapter, I will pursue this line of thought further by investigating ambivalence in relation to free will. In the process, I will develop in more detail some key concepts, such as fairness and retributivism. My main aim here is to show that allowing for ambivalence about free will offers a way of reconciling free will skepticism with compatibilism.

The debate between free will skeptics and compatibilists is at an impasse. Rather than keep looking for conclusive disproof of one side or the other, another way out of the deadlock is to reconcile the two theories. There are, however, two major hurdles that must be overcome for such a project to succeed. First, there is a metaphysical problem: how can one reconcile the skeptic's claim that free will does not exist with the compatibilist's claim that it does? At first glance, the two theories simply appear logically inconsistent. Second, there is a practical problem: even if the two theories can be made metaphysically compatible, it is not clear that metaphysical reconciliation is practically significant. There are concerns by each side that the practical ethical consequences of their opponent's metaphysical contentions are at best negligible. In other words, each side thinks that what the other side calls "free will", even if it can be proven or disproven to exist metaphysically, is not a capacity that ought to have much bearing on actual responsibility practices.

In this chapter, I address both the metaphysical and the practical problem. Metaphysically, a kind of skepticism is compatible with a kind of compatibilism because the two theories target different features of our agency. The compatibilist affirms the existence of the capacity for voluntary action, while the skeptic argues that all our actions, despite many of them being voluntary, are

ultimately subject to luck. Both claims can be true at the same time. The possibility for metaphysical reconciliation has been largely ignored because of the conflation of two ideas associated with skepticism: the idea that the ultimate causes of our *voluntary actions* are traced back to spheres outside of our control (what I will call *ultimate luckism*) with the idea that human beings are not capable of voluntary actions in the first place (what I will call *reductivism*).

Moreover, both the skeptic's and the compatibilist's metaphysical claims have significant ethical implications. Voluntary agency grounds many of the intuitive norms in our responsibility practices, including the often overlooked but crucial norm of negative retributivism, which fixes the scope and degree of permissible harmful treatment. The metaphysical fact of ultimate luck, however, brings up skeptical concerns of fairness about imposing harms or benefits on agents in virtue of their voluntary actions, and hence about negative retributivism. Thus, the skeptical and the compatibilist perspective together entail ambivalence: values will be sacrificed whether we retain or abandon retributive practices.

I cannot emphasize enough that *I will not attempt to prove the metaphysical and practical claims of skepticism or compatibilism here*. Though I will mention some important considerations that speak in favor of both theories, my main purpose is not to add more arguments favoring one side of the debate on free will (this is also why I will not discuss the plausibility of libertarianism). Much ink has already been spilled on this. Instead, I aim to make a *structural point* about the conflict between skepticism and compatibilism. My aim is to demonstrate the possibility of an account of free will that accommodates much of the two deadlocked theories.

I will proceed as follows. I will set my theoretical foundations by surveying some existing attempts at reconciling skepticism and compatibilism—focusing mostly on a strand of theorists developing a framework put forth by Bernard Williams—and by defining several forms of retributivism (§3.2). I will then distinguish between two often conflated ideas associated with free will skepticism: reductivism and ultimate luckism, showing that the latter is metaphysically compatible with a form of compatibilism (§3.3). Having addressed the metaphysical challenge to reconciliation, I will spell out the practical ethical implications of skepticism and compatibilism: acknowledging both entails a number of ambivalences in responsibility where fairness and retributivism come into conflict (§3.4). Finally, I will defend the reconciliatory

project, most notably from worries about the validity of the skeptic's concern with ultimate fairness (§3.5) before concluding (§3.6).

3.2 Theoretical foundations

3.2.1 Attempts at reconciliation

Some attempts to reconcile skepticism with compatibilism are such in name only, as they end up merely in qualified forms of the former (Honderich, 1988; Pereboom, 2017) or the latter (P. F. Strawson, 2003). Another possible view—developed in great detail by Saul Smilansky (2000)—is that skepticism and compatibilism cannot be reconciled even though they both capture correct ethical intuitions. Instead, the two perspectives entail a *contradiction* about the existence of control (a "fundamental dualism"), which is so distressing that it must be hidden by keeping up the illusion that skepticism is false.

A number of authors have come closer to an actual reconciliatory project by following Bernard Williams (1976, 1995a, 2011) in questioning whether the free will problem is a real problem rather than "a *pseudo-problem*; that is, a mere appearance of a problem, and one that is created by the very process of philosophizing" (Fricker, 2022, p. 272; see also Queloz, 2021a; Queloz, 2022; Russell, 2017, 2022; Williams, 1993).

For Williams (1995a), the free will problem deals with three tiers of concepts that must be reconciled with each other: (1) physical terms (cause, effect, determinism, etc.), (2) psychological terms (choice, deliberation, intention, etc.), and (3) ethical terms (responsibility, blame, praise, justice, etc.). Williams (endorsing O'Shaughnessy's (1980) work) maintains that nothing about determinism or a naturalistic view of the universe undermines the psychological terms in tier 2. Human beings are capable of deliberation, intention, etc.—in short, of voluntary agency—and this is enough to ground some form of ethical concepts (tier 3). We should, therefore, reject the assumption driving the free will debate that we need to either prove or disprove the existence of a libertarian level of agency that goes "deeper" than the level of voluntariness around which our responsibility practices are already centered. In this respect, "the will is as free as it needs to be" (Williams, 1995a, p. 19).

Williams's approach, however, does not entail a shallow or "complacent" compatibilism (Russell, 2017, p. 244). Even though we have enough to secure

some form of responsibility, there is still a tension between the psychological (tier 2) and the ethical (tier 3), because more work needs to be done to bring our ethical conceptions in line with a realistic view of human psychology. In particular, it is a mark of the flawed "morality system"—Williams's term for the contingent contemporary Western moral outlook—that the role of luck in morality is suppressed and explained away (Queloz, 2022; Williams, 1981b, 1995c). A truthful account of human agency must depart from traditional (complacent) compatibilism by acknowledging that while we are free and responsible, we are at the same time ultimately subject to luck. The free will problem, from this "critical compatibilist" perspective, is revealed to be "not a (skeptical) *problem* waiting to be solved but a (troubling) human *predicament* that needs to be recognized and acknowledged" (Russell, 2017, p. 244, emphasis in original).

How much critical compatibilism acknowledges the skeptic's concerns is largely left indeterminate by Williams and the authors developing his thought. What does the tragedy of the tragic human predicament consist in? One option is that critical compatibilism entails no changes to our responsibility practices apart from disappointment (or another similar negative attitude like disillusionment) towards the pervasiveness of luck in those practices. The sense of disappointment, however, might be dependent on contingent cultural aspirations; the more we move away from the ideal of a luck-free morality, the less tragic our predicament would seem. Many traditional compatibilists can easily accommodate this attitude of disappointment since it does not concede much to free will skepticism. Alternatively, critical compatibilism could be seen as collapsing into a (qualified) free will skepticism (like that of Derk Pereboom (2014, 2021)) because it allows worries about ultimate luck to shape responsibility practices by completely eliminating their retributive elements (R. H. Wallace, 2019).

As I will show below, neither of these two readings of critical compatibilism do justice to the ethical complexity of skepticism and compatibilism, nor do they offer enough concession for a true middle ground. I propose instead that following through on the trajectory set by Williams results in a reconciliatory view that is significantly distinctive from both compatibilism and skepticism.

3.2.2 Retributivism and compatibilism

Key to this reconciliatory view is specifying the ethical practices that are at issue in the free will debate. Center stage take two retributivist theses:[1]

POSITIVE RETRIBUTIVISM: Wrongdoers deserve proportionate punishment.

NEGATIVE RETRIBUTIVISM: Wrongdoers are liable to proportionate punishment.

A lot hinges on the details of interpreting these principles. First, here and in the rest of the chapter, I shall use "punishment" as a shorthand for all harmful responsibility responses, ranging from the minor harms of blame (e.g., causing emotional distress, ostracizing) to the harshest kinds of institutionalized punishment. Secondly, the difference between "deserve" and "be liable to" can be spelled out in terms of moral reasons. According to positive retributivism, there is a *pro tanto* positive moral reason to proportionately punish a wrongdoer. According to negative retributivism, a moral reason protecting innocents from punishment does not apply to wrongdoers. Notice that negative retributivism does not provide any positive reasons to punish; they need to come from, say, instrumental considerations such as deterrence and moral reform. Thirdly, the same normative logic of positive and negative retributivism can be applied to reward (i.e., beneficial responsibility responses): a commendable agent deserves, or may receive, proportionate reward. We may respectively call these *positive reward-retributivism* and *negative reward-retributivism*.

While some doubt positive retributivism, it is a lot less controversial that negative retributivism is, and should be, at the very core of our responsibility practices. This is because it limits the scope of who can be subject to punishment and how much, even if one believes that the positive reasons for inflicting punishment are instrumental. Without the moral distinction between innocent and guilty of negative retributivism, an instrumentalist system of punishment runs into the well-known problem of being committed to punishing the innocent or punishing out of proportion when that would bring about the best consequences (McCloskey, 1957). Negative retributivism thus permeates our everyday responsibility exchanges: when an agent feels justified in giving the

[1] These definitions build on Mackie (1982) and Alec (2020).

cold shoulder to someone who has wronged her, even if she is doing it with an instrumental purpose in mind, she is probably assuming some form of negative retributivism (Pereboom, 2021, p. 85).

For the rest of the paper, I will focus on those kinds of compatibilism that affirm the existence of a capacity—let's call it *voluntariness*[2]—that can in principle justify retributivism. Regardless of whether these theories spell out the capacity for voluntariness as, e.g., the ability to endorse one's desires (Frankfurt, 1971; Watson, 1975) or to be responsive to reasons (Fischer & Ravizza, 1998; Wolf, 1987, 1994), they share the idea that some aspects of our psychology (tier 2 concepts: deliberation, intention, etc.) are enough to give non-instrumental retributive reasons to treat innocents and wrongdoers differently. I will therefore not discuss instrumentalist forms of compatibilism that justify reward and punishment ultimately by their overall beneficial effects (e.g., Bok, 1998; Smart, 1961; Vargas, 2013).

3.3 Free will skepticism and ultimate fairness

3.3.1 Reductivism

Let us now turn to skepticism. Concerns about ultimate fairness have often been misunderstood because of the conflation of two ideas regarding luck, each one of which can be called a kind of free will skepticism. One is the idea that, *despite being capable of voluntary actions*, we are all ultimately subject to luck; this is *ultimate luckism* and I will come back to it in the next subsection. The other is what I shall call *reductivism*: the idea that determinism or some naturalistic view of the universe entails that human beings are no more capable of voluntary action than other objects in the physical world, and hence cannot properly be called agents in the first place. On Williams's framework, reductivism is the view that physical concepts (tier 1) prove psychological concepts (tier 2) illusory, and thereby also undermine ethical concepts (tier 3).

Free will skepticism has been characterized as reductivism in two of the most influential essays on free will from the 20th century: Thomas Nagel's (1979)

[2] For simplicity, I will talk about voluntariness because many compatibilists of the kind that can justify retributivism agree that one way to distinguish between wrongdoers and innocents is in virtue of their voluntary actions. Some compatibilists, however, believe that people can be held responsible even for nonvoluntary aspects of themselves, such as their attitudes or involuntary failures of attention (Hieronymi, 2008; Smith, 2005, 2015).

"Moral Luck", and P. F. Strawson's (2003) "Freedom and Resentment". These two papers have to a large extent set the framework for the debate that has followed. For Nagel, free will skepticism becomes possible because

> . . .as the external determinants of what someone has done are gradually exposed, in their effect on consequences, character, and choice itself, it becomes gradually clear that actions are events and people things. Eventually nothing remains which can be ascribed to the responsible self, and we are left with nothing but a portion of the larger sequence of events, which can be deplored or celebrated, but not blamed or praised. (Nagel, 1979, p. 37)

As opposed to an internal (subjective) standpoint from which we experience ourselves as subjects capable of voluntary action, taking an external (objective) standpoint "leads to the feeling that we are not agents at all, that we are helpless and not responsible for what we do" (Nagel, 1986, p. 110). Nagel maintains that our ability to see ourselves from both standpoints makes the free will problem irreconcilable.

It is difficult to know exactly how Nagel's intuition that "actions are events and people things" is linked to the observation that our actions are caused by factors outside of our control. I will not pursue the matter in much detail here, though his descriptions suggest that he may have failed to sufficiently distinguish between "causality that runs through the agent and causality that does not" (Williams, 1995a, p. 12). Nagel appears to be imagining that our experiences of agency are mere epiphenomena—that we are observing our actions as if watching TV, like the spectators inside John Malkovich's head. Viewed externally, we appear "helpless", and there is nothing for us to "contribute [to an outcome] as a source, rather than merely as the scene of the outcome" (Nagel, 1986, pp. 113-114). A naturalistic account of the universe thus diffuses responsibility for an action as much as "the information that a particular action was caused by the effects of a drug" (Nagel, 1986, p. 125).[3]

In accord with Nagel, Strawson explains the free will problem as a tension between the subjective and objective standpoint. In a later work, *Skepticism and Naturalism,* Strawson explicitly agrees with Nagel's framing: from an objective

[3] Similar doubts about whether we really causally contribute to our actions have been discussed in response to the experiments performed by Libet and his team, which seemingly prove reductivism (Libet, Gleason, Wright, & Pearl, 1983).

standpoint, we see "people and their doings. . . as natural objects and happenings, occurrences in the course of nature", and "what simply happens in nature may be matter for rejoicing or regret, but not for gratitude or resentment, for moral approval or blame, or for moral self-approval or remorse" (1985, pp. 25-26).

Strawson assumes that if determinism poses a threat to moral responsibility, it must do so by entailing that we should adopt an objective standpoint universally, from which we see any given person as "an object of social policy. . .to be managed or handled or cured or trained" (P. F. Strawson, 2003, p. 79). But contrary to Nagel, Strawson believes that no amount of rational argumentation can threaten viewing the world from the subjective standpoint, as the framework of participant emotions and attitudes "neither calls for, nor permits, an external 'rational' justification" (P. F. Strawson, 2003, p. 91). Free will skepticism is, for Strawson, "idle" (P. F. Strawson, 1985, p. 11). Strawson's view, in contrast to Nagel's, is a form of compatibilism.

Despite their differences, Nagel and Strawson both share a picture of what gives rise to the free will problem. This picture is based on assumptions of 20th-century debates about physicalism (revolving around the work of the Vienna Circle) and sets up the free will debate as a conflict between two standpoints: the compatibilist one, which is concrete, subjective, human, and emotional, and the skeptical one, which is abstract, objective, object-oriented, and dispassionate—what Wilfrid Sellars (1962) has influentially termed respectively the "manifest" and the "scientific" images of humanity.

For many compatibilists, this framing is a non-starter, as they believe that nothing we know about the universe—be it deterministic or indeterministic—reveals that we are incapable of the kind of voluntary action that matters to us. Unlike physical objects, we can form and critically evaluate intentions, and we have persisting commitments and values that form our characters (e.g., Dennett, 1984; Frankfurt, 1971; Watson, 1975). Many of our emotions, and not only the moral ones like indignation and guilt, are apt reactions to these aspects of who we are. Skepticism about voluntary agency—i.e., reductivism—appears to be unfounded, but I will not argue against it here. My modest aim is to show that while there is, in a sense, a problem of free will that essentially arises out of a conflict between two valid considerations, it need not be framed in Nagel and Strawson's way. What they identify as free will skepticism is not the only possible kind.

3.3.2 Ultimate luckism

Even if reductivism is proven wrong, this would not dispel a different worry that follows from a naturalistic account of the universe. Despite our capacity for voluntary agency of a kind that matters to us, our actions are nevertheless ultimately and exhaustively conditioned by factors outside of our control, and this matters to us too. Whether one is a bad person and whether one has had the agential resources to build one's character in an alternative way is ultimately a matter of a complex interplay between one's genes and environment (and, in an indeterministic universe, pure chance). A free will skeptic who follows this path is not skeptical about *voluntary agency.* This view is therefore metaphysically compatible with many contemporary forms of compatibilism. In contrast to reductivism, it can allow for psychological concepts like intentions, choices, deliberation, etc. (tier 2). Skepticism in this form rejects not voluntariness, but rather libertarian attempts to come up with a deeper agential capacity, a kind of contra-causal free will that would conceal the fact that we are all *ultimately subject to luck*. Let's call this form of skepticism *ultimate luckism;* I will assume that it is true for the rest of the paper. The salient feature of reductivism is that agency is rejected; the salient feature of ultimate luckism is that agency is seen as fully conditioned by luck. Ultimate luckism can explain many of the ethical intuitions of contemporary skeptics and is friendly to Williams's project of aligning ethics (tier 3) with what we know about psychology and agency (tier 2).

Why is the universal reign of ultimate luck an ethically important fact? To address this issue, it is helpful to look at the ways in which free will skeptics in the last few decades have elaborated on the moral features of their position in more and more depth, keeping up with a parallel development in compatibilism. One idea that has come out of these developments is that the problem of ultimate luck poses a challenge only to the *retributive* aspects of moral responsibility.[4] Pereboom (2021) captures this line of reasoning by arguing that skepticism targets what he calls "basic desert": the thesis that a person deserves treatment that is experienced as harmful/beneficial by her in virtue of her wrong/right act (performed with sensitivity to its moral status) and

[4] Skeptical theories that explicitly or implicitly challenge retributvism include Cohen and Greene (2006, p. 217), Levy (2011, p. 4), Pereboom (2014, p. 2), Galen Strawson (1994, p. 9), and Waller (2011, pp. 8-ff).

not because of e.g., contractualist or consequentialist considerations (pp. 11-12).

Basic desert, strictly interpreted, is roughly equivalent only to positive retributivism and positive reward-retributivism. As I shall argue below, skeptics are wrong to restrict their focus so, but for now it is worth investigating this formulation for the sake of understanding the ultimate luckist's view. Note also that unlike previous formulations of basic desert (e.g., Pereboom, 2014), Pereboom's most recent definition does not deem blame and praise as inherently irrational. Rather, responsibility practices are threatened by his theory insofar as they are experienced as *harms* or *benefits* by the agent. Pereboom has thus moved further away from a seemingly reductivist view to an ultimate luckist one that allows for forms of blame and praise, along with many emotions directed at voluntary agency, such as disappointment or love.

Now, what makes an ultimate luckist draw the inference from an agent's being subject to ultimate luck to her not deserving retribution? What is doing the normative work to reach this *particular* conclusion: why ultimate luck and why retributivism? If we have voluntariness, a compatibilist might argue, why should we care at all if, after tracing the causal chains of our actions, it turns out that they ultimately go beyond our control?

A clear statement of what justifies the ultimate luckist's inference is Neil Levy's invocation of a *principle of fairness*: "agents do not deserve to be treated differently unless there is a desert-entailing difference between them" (Levy, 2011, p. 9). Strictly speaking, Levy's formulation is formal (as opposed to substantive (Feinberg, 1973, p. 100)) because it does not specify what counts as being "treated differently" and what makes a difference between agents "desert-entailing". For precision, it would be helpful to refer back to my formulation of a more substantive principle of fairness that can still do the normative heavy lifting needed to support skepticism:

> *FAIRNESS:* people should not be treated differently, in terms of burdens and benefits, for what is outside of their control.

I shall use the term "unfair" to denote an act that flouts this substantive principle of fairness, which, however, leaves open the question of whether the act is not morally justified all things considered.

To understand how *FAIRNESS* supports ultimate luckism, we should distinguish between two different levels on which it can operate. On a *surface level*, it is fair to reward two people equally if they contribute equally to an outcome—such as creating a product—and reward them differently in proportion to how much their contributions differ. When it comes to imposing sanctions, again, it seems fair to punish criminals in proportion to their contribution to a crime. If two criminals break the law to the same extent, it would strike us as arbitrary and unfair if we punish one more than the other. So far it seems like we have compatibilist-friendly standards of fairness (Brink & Nelkin, 2013; R. J. Wallace, 1994). Free will skepticism has not entered the picture.

However, this is so only if we remain focused on the surface level, establishing only the extent to which people contribute towards outcomes. The ultimate luckist's point is that instead we should turn our attention to a *deeper level* of human agency. If we follow the causal chains of our actions far enough, we will see that none have their origin in the agent herself, except possibly for such things as chancy neuron firings. The agential structure of a person is ultimately traced back to a combination between genetic makeup and encounters with one's environment. Thus, none of us has ultimately contributed differently—on whatever interpretation of contribution—for becoming who we are. It is as if prior to birth, each of us threw a handful of dice that set out our starting point and our subsequent life-paths in their entirety (including all our voluntary actions). It is worth emphasizing how deep luck goes on this account. It is of course true that in general we have the ability to control and shape who we become. However, from the perspective of ultimate luckism, Daniel Dennett (1984) is wrong to suppose that this saves us from luck, because *how* we shape our agency is also ultimately conditioned by factors outside of our control, such as upbringing and chance encounters (Levy, 2011, p. 199; Russell, 2017, p. 252).

The realization that luck reigns on the ultimate level, coupled with a principle of fairness, yields the conclusion that it is unfair to receive special burdens and benefits in virtue of one's wrong or right actions. No contribution, no retribution. Because the principle of fairness is applied on the level of the ultimate causes of our actions, where ultimate luck holds universally, we may call this line of reasoning a concern for *ultimate fairness*. The grandiloquence of the terms "deep level of agency" and "ultimate fairness" might bring up images of whiteboards in dusty university rooms filled with convoluted metaphysical

formulas that yield abstract ethical conclusions. Yet, even though ultimate luckism is a relatively complex metaphysical picture, its ethical effect—the thought "There but for the grace of God go I"—nevertheless appears to be as immediate and striking as other ethical intuitions.

Distinguishing between the shallow and the deep level explains how compatibilists and skeptics can draw different conclusions from the same principle of fairness. The difference comes from the object of evaluation. On the surface level, it is such things as contribution towards outcomes. On the deeper level, it is contribution towards such things as one's agential structure (and one's abilities to form one's agential structure), which are the *conditions for* making contributions to outcomes (that is, to contributions on the shallow level).

The literature on free will has so far not investigated the possibility that both surface-level and ultimate-level fairness can be valid ethical considerations. Any action can be evaluated on both levels, and there is no reason why one should necessarily eliminate the other. The compatibilist is correct that it can be fair to impose sanctions on a wrongdoer. The ultimate luckist, however, is also correct that it is unfair to do so.

"But why *should* we go this deep?", the compatibilist may ask. One answer is that it simply follows from a consistent application of the principle of fairness. Someone who values fairness can be charged with inconsistency if she restricts its application only to the shallow level of agency. Moreover, many believe that extending concerns about fairness on deeper levels of agency constitutes moral progress. One salient example is the revealing of structural injustices, which arise when certain groups are prevented from, say, acquiring abilities (i.e., conditions for producing outcomes) that are necessary for being professionally competitive (i.e., producing outcomes). Ultimate luckism is a consistent extension of the same kind of reasoning. It is not only the development of professional and life skills, but also the conditions for the very exercise of voluntary agency that are traced back to forces ultimately outside of the individual's control. I will return to some additional worries about ultimate fairness in §3.5.

To conclude: we should distinguish between two kinds of free will skepticism that are often conflated. Though both start from some form of naturalistic worry about our lack of control, the two target different aspects of our agency and lead to different ethical conclusions. Reductivism is skepticism about voluntariness.

It implies that relating to humans as more than mere objects is irrational. All moral and non-moral reactive attitudes (including perhaps even ones like romantic love) would be unfitting, since they would not be targeting real agents at all. Even though I will not argue against it, I have expressed doubts about its truth and will set it aside in what follows. Ultimate luckism affirms voluntary agency but maintains that it is nevertheless ultimately subject to luck. This is the kind of skepticism that I will discuss in the rest of the paper. If ultimate luck is a fact, so is the ultimate unfairness of much of our responsibility practices and acknowledging it cannot leave our ethical life as it is. Though this second kind of skepticism arises out of a realization that we are inescapably continuous with and conditioned by the rest of the world, it does not imply that we are mere objects. Quite the opposite. It is motivated by a concern for treating *human agents* fairly, beings who are capable of voluntary action, but, tragically, lack control over the ultimate forces that shape their voluntariness.

3.4 Reconciliation

3.4.1 Skepticism and compatibilism in practice

Free will skepticism (as ultimate luckism) is not metaphysically at odds with most contemporary forms of compatibilism because it does not deny the existence of voluntariness. But the second hurdle for reconciliation, what I called above the practical problem, is yet to be resolved.

A compatibilist could question whether ultimate luck entails anything of significance to our responsibility practices. For instance, Gary Watson leaves open the question of whether retributivism (by which he seems to mean positive retributivism) is challenged by worries about ultimate luck but points out that it is anyways incompatible with an ideal of universal love (Watson, 2004a, p. 257). It appears that skeptics are not bringing anything new to the table if free will skepticism just collapses into skepticism about positive retributivism, as many compatibilists already reject it for non-skeptical reasons, such as that it is a cruel and irrational practice.

There are, however, a few features of free will skepticism that make it a practically significant and distinctive thesis. First, skepticism has practically different consequences from, say, an ideal of universal love that motivates care for others, because its scope is much broader. Skepticism renders unfair not only responsibility *harms*, but also responsibility *benefits*. Reward-retributivism is just

as ultimately unfair as retributivism about punishment. Further, it renders unfair not only *positive* (reward-)retributivism, but also *negative* (reward-)retributivism. Universal love entails mercy, but it gives no ethical reasons that count against agents receiving welcome or unwelcome treatment solely in virtue of their responsibility.

Second, it is important to draw a distinction in ethical nature and ethical weight between a sense of fairness and other considerations in this context. A libertarian like Kant (2017) might not be swayed by concerns of sympathy or universal love to abandon positive retributivism, but might instead count as a conclusive reason against it the lack of contra-causal free will that would reveal it to be a profoundly unfair practice. Part of what makes Watson's point about universal love so compelling is his detailed description of the real-life criminal Robert Harris's horrific upbringing. Our sympathies are stirred up by seeing Harris who is compellingly presented as a victim of his circumstances (Watson, 2004a, pp. 235-238). However, there are many wrongdoers with a relatively trouble-free upbringing (the "bad apples" (Watson, 2004a, p. 246)). It might be impossible, or at least a lot harder, to see them as *victims* of their environments. For skeptics, the causal story of their actions would still make it ultimately unfair to punish them. The thought "It is a matter of luck that *I* am not *her*" carries a special kind of ethical force. Furthermore, as I mentioned in the previous section, anyone who values fairness but rejects applying it on the deeper level can be charged with inconsistency (and hence perhaps hypocrisy). A concern about ultimate fairness, unlike an ideal of universal love, seems to rationally follow from most compatibilists' existing commitments about fairness in responsibility.

Third, free will skepticism implies that there is a *moral cost* to responsibility practices even if they are justified all things considered. For instance (and as I discussed in §2.4), a system of punishment is justified by its consequences because it is practically necessary to sustain society by deterring bad acts. We cannot live together without wrongdoers suffering some sanction even though they do not ultimately deserve it. There is no way to offset the ultimate unfairness of punishment; compensating criminals for being sanctioned would be obviously self-defeating, as the whole point of deterrence is that one should suffer unwelcome treatment for breaking the law (Smilansky, 2011, 2017). Justice Holmes eloquently puts the moral cost of punishment in the face of free will skepticism:

> If I were having a philosophical talk with a man I was going to have hanged. . . I should say, I don't doubt that your act was inevitable for you but to make it more avoidable by others we propose to sacrifice you to the common good. You may regard yourself as a soldier dying for your country if you like. But the law must keep its promises. (Holmes, 2012, p. 216)

Notice, however, that though the skeptic must recognize the need for a system of punishment, she need not accept *any* system of punishment. Specifically, skeptics are morally committed to a system that, while societally necessary, is also instrumentally good for the wrongdoer herself. To take a somewhat exaggerated example: on a purely consequentialist calculation, it would probably be best for society if a recidivist with low chances of correction (like perhaps Robert Harris) receives capital punishment (as he in fact did). However, from a skeptical perspective, Harris's punishment is ultimately unfair and so there is pressure to impose a kind of punishment on him that, while serving its necessary deterring function, would also be instrumentally good for him. For example, resources might be spent on increasing the chances for his moral correction and reintegration into society.

In sum, though the practical implications of skepticism may at first glance seem minor and technical, on closer inspection they put pressure on major parts of our responsibility practices. Acknowledging the ultimate role of luck in human life does more than reveal our "fragility and vulnerability" (Russell, 2017, p. 260). It also has wide-ranging interpersonal moral effects. There are of course many ethical considerations at play in this context, some of which may sometimes overlap in purpose with the skeptic's ones (like Watson's universal love), and yet they are distinctive reasons with distinctive normative force.

What about the practical significance of compatibilism? The skeptic might claim that the compatibilist's metaphysical affirmation of voluntary agency has no practical import because what really matters is that retributivism is undermined by ultimate luck. There is no good reason, however, to suppose that ethical intuitions grounded in voluntary agency bear no weight. "It is a matter of luck that I am not her" is true, but so is "She did wrong". There is thus room for various important compatibilist intuitions: some compatibilists defend positive retributivism (e.g., McKenna, 2020; S. I. Morse, 2013; Pardo & Patterson, 2013), and many more might agree with it if it is understood as a theory about the harms of (everyday) blame rather than state punishment. More

importantly, compatibilism can justify negative retributivism, without which one would not be able to draw the crucial non-instrumental moral difference between the innocent and the guilty.[5] A skepticism that ignores the negative retributivism grounded in compatibilist voluntary agency would be deeply counterintuitive.

3.5 Ambivalence

There are thus two metaphysical facts about human agency that ground two different sets of conflicting ethical considerations. Retributivist theses are based on our capacity for voluntariness. However, ultimate luck provides special reasons of fairness against receiving harms and benefits in virtue of one's wrong or right actions, which puts pressure on retributivism. A truthful account of the ethics of responsibility should acknowledge ambivalence. Every time someone needs to receive burdens or benefits in virtue of her moral responsibility, we may either impose those burdens or benefits, which would be ultimately unfair, or we may withhold them, which would be a failure to mete out retributive just desserts.

Note that ambivalence incurs not just a psychological cost to the person who must respond to wrongdoing, but also a *moral* one. Something morally bad would happen whatever he does. It would not be enough to resolve this ambivalence to switch between the compatibilist and skeptical perspective at will when that would be to our benefit, as Shaun Nichols (2015, p. 166) suggests. When faced with ambivalence, we are not just facing a psychological conflict that can be resolved by choosing a less troubling perspective. We are faced with a choice that necessarily comes with a moral cost.

Some may suppose that this ambivalence must be suppressed or explained away. Consider for instance Tamlar Sommers, who feels torn between his commitment to free will skepticism on the one hand, and, on the other, the

[5] Pereboom, in building a theory that is ethically acceptable without relying on compatibilism, appeals to the right of self-defence to play the role of negative retributivism. The right to defend oneself, Pereboom claims, meets the challenge of justifying punishment and fixing the scope of who is liable to it (Pereboom, 2021, pp. 86-ff). However, it is not at all clear that the right to self-defence is not just derived from negative retributivism. If that is the case, then Pereboom is sneaking in compatibilist intuitions that are at odds with his official skepticism. Regardless, my point in this chapter is that skeptics do not need to make do without compatibilism; they can instead seek reconciliation.

powerful intuition that someone who wrongly hurts his daughter deserves retributive treatment:

> My considered intuitions are simply inconsistent on this matter. I would like to think that they reflect two aspects of the truth about moral responsibility, but this does not seem to be the case. In the end, then, [free will skepticism] remains the position I endorse upon (tortured) reflection. (Sommers, 2012, p. 201)

For Sommers, retributively punishing the agent who hurt one's daughter feels (retributively) deserved but ultimately unfair, while foregoing punishment feels like the fairer option but also a failure to give what is retributively deserved.

So why would Sommers suppose that he must resolve the ambivalence by choosing one of the intuitions as correct and the other as incorrect? Perhaps he thinks that the two conflicting ethical intuitions stem from two metaphysical pictures that are rationally inconsistent with each other. As I have shown, however, this presupposition is wrong, a remnant from the debate about reductivism. It is indeed irrational to be both a reductivist and a compatibilist, but it is not irrational to accept both voluntariness and ultimate luck as part of one's metaphysics.

Alternatively, perhaps Sommers supposes that even though skeptical and compatibilist metaphysics are compatible, skepticism shows that the compatibilist's *ethical* intuitions are irrational. The proper conclusion to draw from ultimate luckism, however, appears to be that compatibilist distinctions are (ultimately) *unfair* rather than irrational.[6]

3.6 Why not reconcile?

On the reconciliatory account I have proposed, skepticism and compatibilism are metaphysically (and rationally) consistent despite issuing conflicting (and practically significant) ethical considerations that jointly entail ambivalence

[6] On a more fundamental and methodological level, avoiding ambivalence about free will might be part of a problematic general tendency in modern moral philosophy to suppress troubling ethical tensions by issuing binary judgments (Williams, 2011, p. 42). Especially concerning free will, this might be due to an ungrounded "aspiration to *optimism*, in particular to tell a comforting story about the human predicament in respect of moral agency" (Russell, 2017, p. 256, emphasis in original).

about certain retributive aspects of moral responsibility. It appears to be the preferrable way out of the debate's current deadlock, as it recognizes the important ethical and metaphysical truths revealed by each side.

So why would anyone oppose treading the reconciliatory path? First, one might reject the metaphysics of skepticism (as ultimate luckism) or compatibilism. However, to reject the metaphysics of ultimate luck in a satisfactory way, the compatibilist must prove its opposite, namely, the metaphysics of libertarianism. Analogously, to reject the metaphysics of voluntariness, the ultimate luckist must prove the metaphysics of reductivism. Rejecting the opponent's metaphysical picture comes at the price of abandoning one's original position.

Second, one might doubt the ethical conclusions that are drawn from their opponent's metaphysics. Compatibilists may reject that ultimate luck leads to ultimate unfairness and skeptics may reject that voluntariness leads to retributivism. I suspect that breaking apart ultimate luckism from reductivism may help compatibilists to share their opponent's intuitions. If it does not, though, there is not much one can say, except that unless a mistake in one's opponent's theory can be pointed out, reasons of epistemic and ethical humility speak in favor of accepting their intuitions. Furthermore, ambivalence about free will is not a wild idea: there are a number of authors who acknowledge a conflict between skeptical and compatibilist intuitions but do not fully explore its implications (Cuypers, 2013; Nichols, 2015; Russell, 2017), sidestep the issue (Watson, 2004a), or suppress it based on the ungrounded assumption that it is the rational thing to do (Smilansky, 2000; Sommers, 2012).

It is worth dwelling on some special worries about the validity of ultimate fairness, in part because it is the most mysterious and controversial of the intuitions discussed, and in part because such worries may come from Williamsians who are otherwise friendly to the reconciliatory project. Ultimate luckism is a thesis that results from applying standards of fairness on a level that goes beyond the everyday shallow one. Rather than focusing on contributions to an outcome, there is concern about contributions towards the conditions for contributing to an outcome, and about the conditions for those conditions. One might think that something has gone hopelessly wrong with this ethical pursuit. But what could have gone wrong?

Williams charges the "morality system" (the flawed Western modern moral outlook) with distorting ethical life through its aspiration for a completely fair morality (Williams, 1981b, 1995c; 2011, p. 216). One might think that ultimate luckism is just an extension of this problematic project. What is problematic with the morality system, however, is not the concern for ultimate fairness itself, but rather the assumption that it should act as the supreme value that overrides all other (moral and non-moral) values, and that it should exclusively set the normative standards for moral responsibility. I have argued precisely against this assumption by pointing out that ultimate fairness conflicts with retributive considerations. On the reconciliatory account, ultimate fairness is just one ethical force among many.

Another problem with the morality system, as illustrated by Kant's (1998) positing of noumenal freedom, it its tendency of constructing theories that conceal the ultimate unfairness and the troubling ethical ambivalences inherent in responsibility. The way forward is not to abandon the sense of fairness, but rather the opposite: to free ourselves from the illusions that hide its true significance in human ethical life.

A different charge is that the intuition about ultimate fairness comes from setting unreasonably high expectations about responsibility. Assuming that a libertarian concept of free will is incoherent (Nagel, 1986, p. 118; G. Strawson, 1994), or at least impossible for human beings, then for moral responsibility to be ultimately fair one would need to have control over all causal factors conditioning one's agency, such as the sun's shining or the presence of oxygen in the earth's atmosphere. Requiring this kind of control in order to be fairly held morally responsible is a wild criterion, a kind of "metaphysical megalomania" (Fischer, 2006, p. 116). We should stick to the standards of fairness that we *can* have, namely the familiar compatibilist ones. The free will problem, according to this charge, arises out of an absurd wish for ultimate fairness just like the problem of evil arises for those who expect the world to be good (Williams, 1993, pp. 67-68).

This challenge, however, relies on the assumption that we can *choose* where to set the standards of fairness, perhaps because the sense of fairness is a social construct (recall Hume's (2009) view mentioned in §1.2.5). However, fairness is seen by many not as a social instrument but rather as something that we find

intrinsically valuable.[7] To stick with the analogy to the problem of evil: the *philosophical/theological* problem of evil dissolves once one stops expecting the world to be good, but would that make one also stop caring about the evil permeating our world? Many human wishes are impossible or even incoherent, but letting go of them does not assuage the concerns and cares that motivate them in the first place. Abandoning the hope for immortality does not make the atheist stop fearing death or trying to prolong her life for as long as possible. If we could choose what we cared about, there would not be any existential problems to begin with (and philosophers would be largely out of business).

Similarly, the ethical ambivalences that ultimate unfairness introduces to responsibility are not a result of *deciding* that ultimate control should be the criterion of fair treatment. They arise because we seem to care about treating each other fairly, and the sense of fairness commits us to applying it consistently even on the deepest levels of our agency. The direction of explanation between the absurd wish for total control and the concern for fairness goes in the other direction: it is not that we care about fairness because we have a wish for total control. Rather, one may wish for total control because it is the only way to deal with ultimate unfairness. The free will problem can be dissolved *qua* philosophical problem, but by acknowledging, rather than explaining away the human concerns that motivate it.

3.7 Conclusion

A reconciliatory account embraces the metaphysical and ethical claims of skepticism and compatibilism. We are capable of voluntary actions though they are ultimately outside of our control. The former consideration grounds various forms of retributivism, at least some of which lie at the core of our responsibility practices, while the latter reveals retributivism as ultimately unfair. It is an open question how to resolve the resulting ambivalence in any particular situation.

The price of adopting the reconciliatory account is to concede that the ethical intuitions of one's opponent are valid and practically significant. There is no good reason to resist this concession unless one can prove that the opposing theory is mistaken, which has not happened at this stage and does not seem likely to happen. Reconciliation would mark the end of one of the big

[7] Williams too seems to agree that the sense of fairness, unlike the culturally contingent stories and philosophies told about it, is innate and universal (Williams, 1999, p. 248).

longstanding debates in philosophy—we can let go of the divide between skepticism and compatibilism---and make room for a new stage in the study of agency and responsibility.

Chapter 4. Guilt

4.1 Introduction

In Chapters 2 and 3, I focused on the clash between fairness and retributivism that arises in response to the ultimate luck that governs human agency. In this chapter, I will explore a similar moral conflict surrounding another, more ordinary form of luck. Sometimes, an agent may unluckily find herself in a situation that makes it fitting for her to suffer the painful emotion of guilt despite not having done anything wrong. Others might even have reasons to impose guilt on the innocent agent in such circumstances, which creates another ambivalence: the guilt of an innocent may be deserved and yet unfair. To argue for this claim, I will have to develop a novel account of the fittingness of guilt—the Moral Debt Account—that goes against the standard account.

The standard account of guilt (henceforth: the Standard Account) states that it is fitting[1] (and/or deserved)[2] for an agent to feel guilty for an act, attitude, or omission X only if the agent is blameworthy for X. However, there are many cases in which a guilt-like emotion appears to be in some way appropriate even though the agent is not blameworthy. Examples include unintentionally causing harm, as when a lorry driver hits a child through no fault of her own (Jacobson, 2013; Williams, 1981b, p. 236), intentionally performing a wrong act that is justified all things considered (D'Arms & Jacobson, 1994; Hare, 1981, p. 31), hurting someone's feelings in a nonblameworthy way (Shoemaker, 2019), possessing inequitable benefits (Baumeister, Stillwell, & Heatherton, 1994, pp. 247-248; Prinz & Nichols, 2010, p. 134; Zhao & MacKenzie, forthcoming),

[1] E.g., Darwall (2006, p. 71), Wallace (1994, pp. 51-52), Fischer & Ravizza (1998, pp. 5-8), Nussbaum (2004, p. 207), and Shoemaker (2015).

[2] E.g., Carlsson (2017), Clarke (2013, 2016), McKenna (2022), and Duggan (2018). I will discuss the difference between desert and fittingness in §4.6.

and being related to a person or collective that is guilty of serious wrongs, as are the descendants of Nazi war criminals (Telech, 2022).

There are two strategies for defending the Standard Account against these seeming counterexamples (Szigeti, 2015). One is to argue that such agents are not experiencing guilt but rather a non-moral emotion like (agent-)regret. Alternatively, one can show that even though guilt in such circumstances is unfitting, it can nevertheless be appropriate along some other normative dimension(s)—it can be, for example, prudential or morally admirable (Jacobson, 2013; Kamtekar & Nichols, 2019).

Some alternative views, however, hold that guilt can be fitting even if the agent is not blameworthy. One such view states that merely causing a morally bad outcome is sufficient for fitting guilt (G. Taylor, 1985, p. 91). This seems to be too broad. It has counterintuitive implications in cases where one is a mere link in a tragic causal chain, as for instance when one's car is rear-ended and pushed into a pedestrian (D'Arms & Jacobson, 2023, p. 172). More modest alternative views hold that besides blameworthy behaviour, guilt is fitting when an agent possesses inequitable benefits (Prinz & Nichols, 2010, p. 134), fails in her role-responsibilities (Woo, 2023), cannot justify standing in a morally asymmetrical relationship (Zhao & MacKenzie, forthcoming), hurts someone's feelings (Shoemaker, 2019), or commits a personal betrayal (D'Arms & Jacobson, 2023, p. 206). These narrower views overlap with each other only partially; it is an open question whether they can be unified or else how one can adjudicate between them.

The Moral Debt Account of guilt that I will defend in this chapter states that guilt is fitting not only when an agent is blameworthy, but rather, more capaciously, when an agent has a moral debt in virtue of a disrupted moral relationship. The Moral Debt Account can explain the intuitions behind many of the nonstandard views mentioned above and avoids the need for dual-aspect theories on which guilt is fitting *either* when an agent is blameworthy *or* in some other specific circumstances. Furthermore, I will account for some of the conflicting intuitions regarding guilt without blameworthiness by showing that it can be fitting despite being pointless, experienced in epistemically uncertain situations, or undeserved.

This is how I will proceed. I will set my methodological foundations by defining fittingness and outlining the Moral Debt Account of guilt (§4.2) before moving

on to exploring three kinds of cases where guilt over morally innocent behaviour is fitting: guilt over positive inequity (§4.3), guilt in a state of uncertainty (§4.4) and guilt over justified moral costs (§4.5). I will then address worries about desert (§4.6) and conclude (§4.7).

4.2 Fittingness and guilt

When is guilt fitting? To answer this, we must first get clear on what guilt and fittingness are. Let us start with the latter by repeating and elaborating a bit on my characterization of fittingness in §1.2.4. Recall that fittingness is a normative concept that applies to the relation between a person's attitude or emotion and an object. An emotion is fitting vis-à-vis its object if it correctly *appraises* it.[3] Irrespective of whether this appraisal is spelled out in a cognitivist or non-cognitivist way, much of our intuitions about the gloss of a given emotion's distinctive appraisal depends on the emotion's typical elicitors, its phenomenology, and its motivational profile (D'Arms & Jacobson, 2023, p. 142). Regarding emotional motivation, it is useful to distinguish between an emotion's *goal* and its *action tendencies* (D'Arms & Jacobson, 2023, p. 105). The goal is what satisfies the emotion, while the action tendencies are the direct and urgent actions that the emotion motivates one to perform. The goal of fear is threat avoidance, while its action tendencies include hiding, fleeing, and freezing. The goal and action tendencies can come into conflict; sometimes the best way to avoid a threat is to confront it rather than to hide, flee, or freeze.

In short, specifying the fitting objects of guilt requires an accurate gloss of its appraisal, and this in turn depends on having a clear view of guilt's elicitors, phenomenology, goal, and action tendencies.

What is guilt, then? The word "guilt" can refer either to culpability or to a distinct negative self-directed emotion, which is unsurprising as the two often go together. "I feel guilty for X" usually means that I judge myself blameworthy

[3] Let me mention again that two prominent theories of fittingness are representationalism (Milona, 2016; Roberts, 2003; Tappolet, 2016) and attitudinalism (Deonna & Teroni, 2012, 2015, 2022; Mitchell, 2021; Mulligan, 2007; Müller, 2017). According to the former, an emotion or attitude is fitting if it correctly *represents* its object; according to the latter, an emotion or attitude is fitting if it correctly *engages with* its object. Both, however, define fittingness in terms of correct appraisal.

for X and that I have (the disposition to feel) the emotion of guilt about X.[4] In fact, many philosophers believe that the emotion of guilt necessarily involves the appraisal that one is blameworthy; to have a bout of guilt just is a form of self-blame (e.g., Lamb, 1983, p. 340; Nussbaum, 2004, p. 207; Rawls, 1971, p. 445; Sussman, 2018, p. 792). On this picture, guilt cannot be fitting if the agent is not blameworthy. Since my purpose in this paper is to argue that guilt can be fitting even if the agent is innocent, I will focus on the *emotion* of guilt and, in the next section (§3.2), show that it does not necessarily involve appraising oneself as blameworthy.

Here is a paradigmatic case of guilt to serve as the basis for an initial characterization:

> ***Reckless driver.*** Damon takes his friend Vicky out for a ride, but he intentionally omits telling her that he does not have a driver's license. Because he is an inexperienced driver, Damon crashes into a lamppost and both he and Vicky are severely hurt. Damon feels guilty.

Damon's guilt feels painful and motivates him to perform certain actions and to reach a specific goal. The goal of guilt can be glossed as *relationship reparation* (D'Arms & Jacobson, 2022, p. 21; Shoemaker, 2019, p. 143). Let us flesh this out. Guilt is not aimed at merely repairing a bad state of affairs. Damon might be motivated to repair the harm that he has caused Vicky by calling an ambulance out of *sympathy* for her. While sympathy can be expected from bystanders, it is woefully insufficient for Damon. If others' well-being was the only thing that we cared about in such situations, human beings would not need more than emotions like sympathy, and guilt would be functionally superfluous. Damon has damaged more than Vicky's health. His morally wrong—and blameworthy—act has led to the disruption of his moral relationship with the moral community and especially with Vicky. By a *disrupted moral relationship,* I mean that the moral relationship is in an undesirable state. The persons involved need not recognize that the relationship is disrupted for it to be such. That is to say, even if Vicky happens to not care about Damon's wrongdoing, the two nevertheless *ought to* not continue as if nothing morally disruptive had happened. Some responses are called for. Specifically, in virtue of the disrupted relationship, Damon has incurred a *moral debt*, by which I mean that he ought

[4] Because of these two close meanings of the word "guilt", it is difficult to express in ordinary language the state of experiencing the emotion of guilt without judging oneself blameworthy, as in the cases that I will discuss below. We sometimes use phrases like "I feel bad for X".

to perform certain actions in order to restore the moral status quo.[5] It is up to him to make amends.

Based on this description, we can gloss the appraisal of guilt as *having a moral debt in virtue of a disrupted moral relationship* and its goal as *moral relationship repair*. Much of the time guilt targets the disrupted relationship with someone in particular; people tend to punish themselves out of guilt more when in the presence of the particular person whom they have wronged (Nelissen, 2012).[6] Perhaps in time others will think that Damon has done enough to atone, but it would be very difficult for Damon to overcome his guilt if *Vicky* never stops holding the incident against him.

How does guilt help Damon repay his moral debt to Vicky? Vicky has suffered physically and emotionally because of Damon, and she would also be correct to judge that he has treated her with insufficient regard. These moral ruptures would be suitably addressed by the action tendencies of guilt, which can be glossed as *compensation* and *apology* (see e.g., D'Arms & Jacobson, 2022, p. 21; Portmore, 2022, pp. 57-59).[7] (Framing the latter action tendency as apology, however, is biased towards the Standard Account because it is derived from instances in which the agent is blameworthy. As the cases discussed in the following sections will show, apology is a particular form of the more general action tendency of guilt, namely *expressing regard*: the guilty agent wants to manifest to the other that she respects and cares about her and their relationship.)

Besides motivating compensation and expressing regard, the very conspicuous presence of guilt is a signal of a kind of *submission*. The guilty person "hangs

[5] We may speak also of moral debts that are not incurred in virtue of a disrupted relationship. For instance, I may owe someone to fulfil my promise to him, but this is in virtue of my making a promise. Our relationship is not in an undesirable state. These kinds of debts, on my account, do not merit guilt.

[6] One might also have a moral debt that disrupts one's relationship with oneself, which would explain why we feel guilty for cheating on our diets.

[7] Is self-punishment an action tendency of guilt? Some doubt it (Baumeister et al., 1994, p. 256), and while there is empirical evidence that people seek punishment when feeling guilty (e.g., Inbar, Pizarro, Gilovich, & Ariely, 2013; Nelissen, 2012; Watanabe & Ohtsubo, 2012), it is possible that they do so merely as a means of repair, for instance to appease others' resentment. Perhaps if the person who has been wronged would be happy with compensation and apology without punishment, then the guilty agent would not spontaneously feel motivated to punish herself. More research is needed to establish whether self-punishment is an action tendency of guilt or is rather on occasion derived from the more general motivation to repair.

[her] head" (Velleman, 2003); her body feels limp. Guilty Damon is willing to put himself in the power of Vicky to the extent necessary to make up for his wrongdoing; he submits in readiness to repair. The message that Damon's guilt signals to others is the same as its appraisal: "I have a moral debt to Vicky". We can better appreciate the signalling role of guilt in responsibility exchanges if we imagine that Damon did not feel guilty in ***Reckless driver***. Without the readiness to repair that is communicated by the submissiveness of genuine guilt, Vicky would have been licensed to have serious doubts whether Damon was someone whom she could trust to respect her interests. There is evidence that forgiveness usually occurs when there has been a show of guilt on the part of the wrongdoer (for an overview, see Dill & Darwall, 2014, p. 40). If Damon were to go through the motions of apology and compensation without guilt, one could easily think that he was not aiming to repair the moral relationship for the right reasons but was instead driven by some egotistical calculations.

Before moving on to investigating cases of guilt without blameworthiness, we need a method for distinguishing guilt from other closely related emotions, such as shame, (agent-)regret, and remorse. I will assume that guilt is a pancultural *natural emotion*: a distinct psychological kind given with human nature (D'Arms & Jacobson, 2023, p. 106).[8]

One way to type-identify emotions is through their constitutive thoughts (e.g., Foot, 1978; Nussbaum, 2001). This method has certain limitations, however. First, assuming that emotions even have constitutive thoughts, it is very often unclear exactly what the constitutive thought of a distinct emotion is. Second, type-identifying emotions by their constitutive thoughts risks obscuring the division between natural emotions; it fails to carve emotions at their psychological joints (D'Arms & Jacobson, 2023, p. 96). As it will become clear shortly, this is especially problematic when it comes to distinguishing guilt from agent-regret and remorse. Instead of the constitutive thought method, I will adopt a motivational theory of the emotions, according to which emotions can be distinguished from each other based on their distinct motivations (D'Arms & Jacobson, 2023, Chapter 6). Differences between emotions in phenomenology and constitutive thoughts are difficult to discern, while

[8] Not everyone accepts that guilt is a natural emotion. For instance, Velleman (2003) conceives of guilt as a form of anxiety, and Nichols and Prinz (2010, p. 136) suppose that guilt is a combination between sorrow and anxiety. I discuss why guilt is different from sorrow and anxiety below.

emotionally motivated behaviour can be observed more easily. The test for whether a given emotional episode is one of guilt, then, will be determined by whether the agent shows signs of submission and is motivated to compensate and express her regard.

Let us see what the motivational theory yields when it comes to guilt vis-à-vis other emotions. A closely related self-directed emotion is *agential regret* (not to be confused with *agent-regret*!). This is the kind of regret a person feels towards his own rational errors (which can be but are not necessarily moral) and it motivates him to undo them and change policy for the future (Jacobson, 2013, p. 104). When I stupidly invest in cryptocurrency and lose a lot of money, I will regret my action, but I will not feel submissive and motivated to compensate and express my regard.

As I mentioned above (§4.1), a standard strategy of the Standard Account to deal with situations that are problematic for the theory is to argue that they merit *agent-regret* (not to be confused with *agential regret!*) rather than guilt. Agent-regret—an emotion term coined by Bernard Williams—is supposed to be constituted by a thought like "I wish that I had done otherwise" (Williams, 1981b, p. 27). However, agent-regret spelled out in this way does not appear to have any distinct motivations; it does not add anything that cannot be explained by guilt and agential regret. Williams himself says that agent-regret is accompanied by the desire to compensate (Williams, 1981b, p. 28)—which is among the motivations of guilt—and furthermore that agent-regret "can be psychologically and structurally a manifestation of guilt" (Williams, 1993, p. 93). On the motivational theory, unless we can identify a distinct motivation for agent-regret, we should suppose that it is not a distinct natural emotion but collapses into either guilt or agential regret.

The same can be said about remorse. Remorse has been conceptualized in quite different ways. For example, remorse focuses on deeds, while guilt targets the self (G. Taylor, 1985); or instead, remorse is a fitting response to involuntary wrongdoing, in contrast to guilt, which targets voluntary wrongdoing (Mason, 2019, p. 190). From the point of view of the motivational theory of emotion, these accounts of remorse would fail to establish it as a distinct natural emotion unless a distinctive motivation is identified. Otherwise, just like with agent-regret, remorse would be referring to guilt or agential regret in a particular set of circumstances. Perhaps remorse is a useful concept nonetheless, but not one that carves out a natural emotion.

Guilt is different from shame, though both may be felt in response to moral wrongdoing. Shame targets something that reflects badly on oneself (D'Arms & Jacobson, 2023, p. 149) and motivates concealment rather than compensation and expressions of regard (Deonna & Teroni, 2008). Guilt is also not a form of anxiety that one is facing the prospect of blame, punishment, or the loss of trust (Velleman, 2003). Not only does anxiety lack the distinctive motivations of guilt, but also a person would plausibly feel guilty even if absolutely sure that she will not be punished, blamed, or lose another's trust. Lastly, sadness (or sorrow), in contrast to guilt, is not self-directed. Vicky's brother may be sad that she was hurt in the accident, but only Damon would also feel guilty.

To sum up, on the Moral Debt Account, guilt is a painful self-directed natural emotion, which appraises the agent as having a moral debt in virtue of a disrupted moral relationship, and it has as its goal repairing this relationship. Guilt signals readiness to repair through submission and motivates compensation and expressing one's regard. Proponents of the Standard Account would most likely agree with me that agents like Damon possess a moral debt in virtue of a morally disrupted relationship. They would deny, however, that moral debts besides those produced by blameworthy behaviour make guilt fitting. My strategy to prove otherwise will be to discuss three kinds of cases of morally innocent agents who feel guilty for their moral debts and argue that their emotion satisfies the conditions for fittingness. With the methodological groundwork in place, we can now turn to a first class of such cases: guilt over positive inequity.

4.3 Guilt over positive inequity

Positive inequity refers to a situation in which a person possesses inequitable benefits in relation to someone else. Let us begin with two cases in which an agent, through no fault of her own, is in a state of positive inequity.

> ***Inheritance.*** Tyler's aunt wills her cottage to him and her porcelain clown figurines to his sister Katerina. Tyler and Katerina were equally close to their aunt, and neither is wealthier. Tyler feels guilty over unfairly receiving a lot more than his sister.

> ***Survivor***. Bonnie is the only survivor of a tragic airplane accident. She feels guilty that due to pure luck, she survived while others did not.[9]

Tyler and Bonnie, though they are not blameworthy, undeservedly fare better than, respectively, Katerina and the other passengers. Take ***Inheritance*** first. There is nothing mysterious about Tyler's guilt if one agrees that fairness is a non-instrumental human concern that makes us averse to social inequity (see §1.2.5). Human emotions seem to track (or even constitute) this aversion. Katerina would be correct to feel righteously envious[10] of Tyler because he has benefitted unfairly at her expense. Their aunt's will has disrupted their relationship and created a moral debt for Tyler: to repair the relationship with his sister, he is morally expected to restore equity through his words, actions, and feelings.

Tyler's guilt is tailored for this task. His submissiveness signals that he feels like he should set the unfair situation right. He might be motivated to express his regard for his sister by assuring her that she is no less deserving of their aunt's legacy than him. Furthermore, he might want to compensate Katerina for the inequity by, for instance, offering to split ownership of the cottage. The signalling and motivational features of guilt are perfectly suited to appease Katerina's righteous envy and restore equity.

Some might suppose that Tyler is not feeling guilty for having a moral debt but is instead experiencing *anticipatory guilt* (Kerr, 2019), i.e., guilt evoked by his entertaining the possibility that he will fail to set things right. However, anticipatory guilt is usually experienced in momentary bouts, and not that often when the agent is reasonably sure that she will not fail in her duties. One is not troubled by much guilt just because one might not pick up one's sick friend from the hospital. Of course she will; no need to dwell on it. In contrast, according to a wealth of empirical data, many people experience guilt over long periods of time and testify that its object is *possessing* inequitable benefits (for an overview, see Baumeister et al., 1994, pp. 247-248; Zhao & MacKenzie, forthcoming).

[9] Both cases are taken from Zhao and MacKenzie (forthcoming)

[10] Righteous envy can be called resentment in ordinary language (Velleman, 2003, p. 246), which speaks in favour of the intuition that it targets the disruption of a *moral* relationship.

Other doubts about the fittingness of guilt over positive inequity can stem from cases of survivor's guilt, such as ***Survivor***, where a person luckily survives while others perish. Many philosophers resist the idea that survivor's guilt is fitting (Metz, 2018; cf. Velleman, 2003; Zhao & MacKenzie, forthcoming). From the lens of the Moral Debt Account, whether Bonnie's guilt is fitting depends on whether her survival disrupts her moral relationship with the other passengers in a way that creates a moral debt. To sidestep difficulties with debts to the dead, let us suppose that rather than dying, all other passengers were severely hurt. Would Bonnie have owed them anything, like Tyler owes Katerina to restore moral balance? Intuitions here may vary. I will explore the perhaps more widely shared intuition that Bonnie's good fortune creates no moral debts. It would be exceptionally kind of her to contribute to the payment of the other passengers' hospital bills, but she does not *owe* it to them.

What can account for this difference between ***Inheritance*** and ***Survivor***? I cannot do justice to the extensive literature on the philosophy of fairness here but can offer a sketch of an answer. One plausible explanation is that Tyler is benefitting from his aunt's actions, whereas Bonnie's benefits are gifts of fortune. This seems to be a morally relevant difference. Similar to how it is fitting to resent the arsonist's actions but not the lightning strike that causes a fire—though both create morally unfortunate outcomes—so it is fitting to resent unjust distributions produced by faulty human agency but not natural ones. One can agree with luck egalitarians that there are moral reasons to eliminate both kinds of inequity (e.g., G. A. Cohen, 2000; Dworkin, 2000; Rawls, 1971; Temkin, 2017) but still maintain that only the former makes the beneficiary in particular incur a debt. We ought not to undeservedly benefit from unjust distributive systems (like racist governments) but we may undeservedly benefit from good fortune. Is one *morally required* to share the briefcase of money that they luckily find on the street? Those who answer negatively can allow that, unlike Tyler, Bonnie has no moral debt. If this explanation is agreeable, Bonnie's guilt is unfitting; she is falsely intuiting that her undeserved good fortune is a result of someone's agency. Only unjust distributions produced by faulty human agency create the conditions for moral debts and hence for fitting guilt in those who benefit from them.

To sum up, guilt over non-natural positive inequity is widely experienced, functional, spontaneous, and not based on any factual errors. Everything about it points to its being fitting. An important result from this conclusion is that guilt does not necessarily involve appraising oneself as blameworthy; guilt is not

a form of self-blame. Rather, guilt is more broadly triggered in situations where a person has a moral debt in virtue of a disrupted moral relationship.

One last problem with guilt over positive inequity has to do with cases like the following:

> ***Tenure.*** Jenna, an accomplished academic, lands a coveted tenure-track job. She knows that she deserves it but feels guilty that the other hundred or so candidates will have to be rejected.

According to Zhao and MacKenzie (forthcoming), unlike in ***Inheritance*** and ***Survivor***, there is no problem of fairness in ***Tenure*** because Jenna deserves the job. Jenna's guilt should be explained instead as what Zhao and MacKenzie call *solidarity guilt*, which arises out of a requirement to share the fates with other group members. In virtue of her group membership, the agent feels that she owes it to other group members to eliminate certain disparities between them.

I disagree with Zhao and MacKenzie that there is a need to posit such a thing as solidarity guilt. ***Tenure*** is simply a case of positive inequity, just like ***Inheritance*** and ***Survivor***. A state of affairs in which one person, even if she is the most deserving, gets all the benefits, while everyone else gets nothing is not equitable. This is because the distribution of benefits does not correspond to what all the candidates deserve. Presumably, many of the other candidates' efforts and talents merit benefits too, and it is unfair that they will be declined employment (though this distribution of goods may be, in that time and place, morally justified all things considered). Jenna's emotions are luckily not victim to false consciousness. Her guilt is an acknowledgment that she is benefitting from an inequitable distributive system. To make the point even clearer, imagine that Jenna was only competing with two other candidates who were obviously unqualified and just did not deserve the job. She might indeed have experienced some sympathy for them, but guilt would have been out of place.

4.4 Guilt in a state of uncertainty

Positive inequity is only one kind of situation that merits guilt without the agent being blameworthy. Another one, which I will investigate in this section, is when the agent incurs a moral debt through falling into a particular state of uncertainty about her blameworthiness. Here, I will take as the central example Bernard Williams' (1981b) widely discussed lorry driver case.

Lorry driver. A lorry driver hits a child through no fault of his own. He feels guilty.

As I argued above (§4.2), though Williams posits agent-regret to capture the lorry driver's feelings, there are serious doubts about whether there is in fact such a distinct natural emotion. The lorry driver feels guilty, as many real-life drivers in similar circumstances (Anderson, Kamtekar, Nichols, & Pizarro, 2021; Wojtowicz, 2022). His guilt, furthermore, is not about positive inequity; it would be strange if his primary worry is that he is inequitably faring better than the child.[11]

Lorry driver has stirred much debate because it gives rise to conflicting intuitions. On the one hand, since he is not blameworthy, there are reasons why he should overcome his guilt: "[d]oubtless, and rightly, people will try, in comforting him, to move the driver from this state of feeling. . . to something more like the place of a spectator" (Williams, 1981b, p. 28). On the other hand, it is expected of the lorry driver to experience guilt, as "some doubt would be felt about a driver who too blandly or readily moved to that position [i.e, the place of a spectator]" (Williams, 1981b, p. 28). What sort of shady person would he be if he does not feel guilty, at least initially? A satisfactory account of the case must explain both intuitions (Kamtekar & Nichols, 2019, p. 183).

Defenders of the Standard Account can meet this challenge in a tidy way by claiming that the lorry driver's guilt is unfitting but morally admirable. There are two ways to make this case. One is to argue that though the lorry driver's guilt misfires in the present circumstances, this serves as a signal for others that it will fire correctly when the agent is in fact blameworthy. The lorry driver's irrational guilt is a sign that he is a morally decent agent since he has a functioning guilt mechanism (though it is in need of some attuning (Kamtekar & Nichols, 2019)).

A different strategy, pursued by Daniel Jacobson (2013), is to highlight that ***Lorry driver*** is a stylized case, which conceals that on a natural reading of the case, the lorry driver will not be sure whether he is blameworthy immediately following the accident. Consider a contrast case:

[11] Zhao and MacKenzie's (forthcoming) moral imbalance account can therefore not account for cases of the kind discussed in this and the next section. Not all fitting guilt without blameworthiness is due to a moral imbalance if this is spelled out, as the authors seem to do, in terms of a distribution of burdens or benefits.

> ***Causal link***. Caroline stops her car on a red light. Another driver rear-ends her, causing Caroline's car to fly forward and hit a crossing pedestrian, breaking her foot. She does not feel guilty.

Let us assume with D'Arms and Jacobson (2023, p. 172) that Caroline would not and should not feel guilty. This serves to show that a mere causal contribution to a morally bad outcome is not sufficient for fitting guilt. The relevant difference between Caroline and the lorry driver appears to be that the latter has much more ground for doubt about whether he is at fault. Was he driving negligently, with an unsatisfactory degree of attention? Was there a speed limit sign that he had missed earlier down the road? Has he checked his brakes recently? Since he cannot be immediately sure about whether he is at fault for hitting the child, erring on the side of caution and presuming blameworthiness—and consequently feeling guilty—is the morally admirable thing to do.[12]

I will argue that pursuing Jacobson's point about uncertainty further, along with appreciating guilt's function to signal readiness to repair a relationship in a dynamic responsibility conversation, leads to the conclusion that the lorry driver's guilt is morally admirable *and* fitting.

In many real-life cases, including ones like ***Lorry driver***, the agent cannot find out whether she is blameworthy, and if so to what degree, *only by herself*, because this depends on a socially set standard. This is especially true in cases of potential negligence, since negligence is defined in terms of a failure to exercise reasonable care, and "reasonable" is open to interpretation. Perhaps the lorry driver could have avoided the incident if he had above-average reflexes, attention, and skills, but ought he have met this high bar? In any case, it would be arrogant of the lorry driver to presume that it is exclusively up to him to decide whether he is to blame. When there is room for doubt, the decent agent submits to the judgment of others, and especially to the judgment of the people directly affected; in this case, the child and her parents.[13] He rationally cannot but be uncertain about his blameworthiness.

---

[12] For similar views, see Enoch (2012) and Mason (2019, Chapter 8). My account, as will become clear shortly, differs from these views in that I do not claim that the lorry driver ought to take on blameworthiness or must act as if blameworthy because of his uncertainty.

[13] The lorry driver need not accept their judgment without question. If he is unjustly blamed way out of proportion, he may feel angry rather than guilty.

For a limited time after the accident, the lorry driver thus stands in a morally disrupted relationship with the child and her family, as he is now possibly culpable for their misfortune. Being in this ambiguous state comes with certain responsibilities. For one, he owes the family a commitment to submit to their judgment (provided it is reasonable) and to make amends if they judge him blameworthy. Guilt, prompted by this moral debt, finds its place here.[14] By emotionally manifesting a readiness to repair, the lorry driver acknowledges that he is putting himself in the power of others—even if he is judged blameless in the end. Further, expressing his regard (e.g., visiting the child in the hospital), shows that he values the child's well-being and is willing to take responsibility if it turns out that he is at fault.

What would the parents of the child think if the lorry driver—despite being in a state of possible culpability—had reacted without guilt and had from the start felt detached from the harm he had caused? Imagine that like Caroline, he had felt, for example, sorrowful but not guilty. The family would have been entitled to think that he was arrogantly refusing to commit to being judged and repairing the harm. The emotion of guilt, therefore, signals the commitment to repair that is needed to begin a dynamic responsibility exchange in the right way. It is part of the responsibility conversation from the start, rather than, as proponents of the Standard Account hold, coming at its end as the required response from the wrongdoer in case she turns out to be blameworthy (e.g., McKenna, 2022). When it comes to emotions, it is guilty until proven innocent.

What about the observation that others will try to bring the lorry driver out of his guilt? According to my interpretation of events, when others are talking him out of his guilt, they are not *correcting a mistake*, but are rather *declaring their moral judgment*: they are bringing him out of his state of uncertainty by letting him know that he is not blameworthy. They are communicating that he should not feel guilty *anymore*, similar to how a wrongdoer who has done enough to atone is reassured that he is free to let go of his guilt. They are saying, "You should not feel guilty" but not, "You should not have felt guilty". Once the lorry

[14] To merit guilt, the agent's uncertainty must be about whether her causal contribution to the bad outcome makes her blameworthy and not about whether she causally contributed to the outcome in the first place. If the lorry driver, due to shock, was unsure of whether he even had been driving the lorry at the time of the accident, he would have been primarily confused and anxious but not guilty.

driver is judged that he is blameless, the goal of guilt—repairing the disrupted moral relationship—is reached.

Here are two possible objections to my account. First, one could point out that guilt appraises the agent as blameworthy, which necessarily means that the lorry driver is making a mistake. Rather than trade intuitions about how the lorry driver is appraising events, we can address this issue by relying on the results from the previous section: agents appear to fittingly feel guilty over non-natural positive inequity despite knowing full well that they are not to blame. Guilt's appraisal, therefore, does not necessarily include blameworthiness. The guilty agent feels instead that she has a moral debt, and this is precisely what the lorry driver seems to feel too: he owes the child and her parents a commitment to submit to their judgments and possible demands for compensation.

A second objection has to do with the normative concept of fittingness. It is conceptually built into fittingness that it is a separate dimension of evaluation from epistemic warrant (Echeverri, 2019): fearing the grizzly bear is fitting even if the agent is perfectly justified in believing that it is an abnormally big raccoon or a plushie. One might conclude, then, that a test for whether the lorry driver's guilt is fitting is to consider the counterfactual situation in which he has full knowledge of whether he is blameworthy. If it is unfitting for the lorry driver to feel guilty knowing, like Caroline, that he is blameless, then guilt is unfitting in his case regardless of his actual epistemic state.

However, the fact that fittingness is separate from epistemic warrant does not entail that the fitting object of an attitude is wholly independent from the agent's epistemic relation to it. Otherwise, we would not be able to make sense of attitudes like surprise and curiosity. Plausibly, surprise appraises its object as something like "unexpected and significant" (Baras & Na'Aman, 2022), and curiosity as something like "interesting but not completely known". Thus, a fitting attitude's appraisal can involve epistemic uncertainty. (Still, even for curiosity and surprise, fittingness is separate from epistemic warrant: an agent can be unfittingly curious about the boring book on his bedside table, falsely but justifiably believing that it is the newly published novel by his favourite author.)

It is possible, then, that guilt is among those emotions whose appraisal may include epistemic uncertainty. Asking whether the lorry driver would have felt guilty if he had known that he was not blameworthy would be like asking

whether you would have been curious about X if you had known everything about it already. It does not tell us about its fittingness.

In short, guilt in cases like ***Lorry driver*** is widely experienced, functional, spontaneous, expected of the agent, and the concept of fittingness allows for his epistemic uncertainty. Still, it is not easy to conclusively arbitrate between the view that guilt in such circumstances is morally admirable and fitting and the view that it is morally admirable but unfitting. One way to approach the problem is to ask: Would the lorry driver and others involved—such as the child's family—feel conflicted about his guilt following the accident once he is judged blameless? Would he and others think that his emotion was irrational despite its being a sign of virtue and moral decency? My intuition is that there would be no such psychological conflict. Feeling guilty after hitting a child on the road would just be perceived as a fully appropriate feeling, with no irrationality involved. The lorry driver's guilt while awaiting judgment is the natural human emotional sign that the child and her parents are owed a respectful commitment to repairing the harm.

4.5 Guilt over justified moral costs

An innocent agent may also fittingly feel guilty when she all-things-considered justifiably harms or wrongs another. Consider the following cases:

> ***Agamemnon.*** Agamemnon sacrifices his daughter Iphigenia to appease the gods and fulfil the duty he has to his countrymen. Though he believes himself morally justified in his horrendous act, he feels guilty.
>
> ***Unrequited love.*** With time, Elena has developed as a person and has thereby grown apart from her partner Stefan. She eventually discovers that she no longer loves him and that they need to split up for her to be happy. Stefan is heartbroken and though both agree that Elena is not blameworthy for the breakup, she feels guilty.[15]
>
> ***Sophie.*** Sophie and her two children are prisoners in a Nazi concentration camp. Sophie's captors make her choose which one of her two children is to be killed; if she refuses to make a decision, they threaten to kill both. Sophie chooses the younger sibling and feels profound guilt even though

[15] This case is inspired by empirical studies of guilt in instances of unrequited love (Baumeister, Wotman, & Stillwell, 1993).

she knows that there was nothing she or anyone else in her situation could have done better.[16]

Let us stipulate that Agamemnon, Elena, and Sophie are not blameworthy and know this. (If that is not believable in ***Agamemnon***, you may substitute the original case with one in which, for example, Iphigenia's sacrifice is necessary to save many lives.) The guilt of these agents, therefore, cannot be explained by uncertainty. Nor can their guilt be made sense of by positive inequity: the ones who suffer from the agents' actions would not plausibly feel righteous envy towards them. Stefan's concerns, for example, have nothing to do with Elena's undeservedly faring better than him; he is simply hurt and disconsolate.

The key to accounting for these cases is the intuition that under certain conditions others are owed compensation when we intentionally harm or wrong them, even if this harm or wrong is morally justified (i.e., when there are moral costs to our justified actions). Though Hollywood films usually omit this part of the story, a detective who commandeers a citizen's car and wrecks it in pursuit of a dangerous criminal owes recompense. Furthermore, financial restitution aside, it is sometimes more important to assure others who foreseeably bear costs as a result of our actions that their interests are important and should not be sacrificed easily.

Harming or wronging others—even if it is morally justified all things considered—disrupts one's moral relationship with them because they are treated as people whose interests can be sacrificed.[17] The agent owes a moral debt in virtue of this relationship rupture: unless compensated, others will remain in an unjust state of inferiority in relation to the agent. This occasions guilt, which signals readiness to repair through submission and motivates one to compensate for the damages inflicted and express one's concern for others' interests. The agents' intentional actions in ***Agamemnon***, ***Unrequited love***, and

[16] ***Agamemnon*** and ***Sophie*** are respectively based on William Styron's (1979) *Sophie's Choice* and Euripedes's *Iphigenia in Aulis.* The two are often discussed as examples of moral dilemmas: situations in which the agent is blameworthy whatever she does (Greenspan, 1983b; Holbo, 2002; Stocker, 1971; Tessman, 2015; Williams, 1973a). As it will become clear shortly, I propose to view them instead as cases in which it is fitting for the agents to feel guilty despite not being blameworthy.

[17] On a Kantian moral theory, sacrificing someone's interests for another's sake is impermissible, and hence Sophie and Agamemnon are blameworthy for their actions. An interesting question, which I will not pursue here, is whether a Kantian who refuses to act in a situation like ***Sophie*** or ***Agamemnon*** would have grounds for fitting guilt.

Sophie carry the meaning "Your interests will be sacrificed", and guilt adds, "but I am indebted to you for doing so".

In ***Agamemnon,*** the moral debt takes dreadful proportions. Death creates a situation in which Iphigenia cannot be compensated. The Greek army, through the agency of Agamemnon, has incurred a moral debt that cannot be repaid, and his guilt will thereby never reach its goal of repairing the disrupted relationship. Not being able to reach the goal of guilt might mean that Agamemnon would be forever stuck with and tormented by it. The stuff of tragedy. Note, however, that the impossibility to reach guilt's goal does not make it unfitting any more than the impossibility of outrunning the grizzly bear makes fearing it unfitting. Agamemnon's guilt may be irrational in some sense—it is to some extent futile—but fitting nonetheless. Though Agamemnon will never restore the relationship with his daughter, his guilt is a fitting acknowledgment of his share in the moral debt to her.

The harm caused by Elena's decision to leave Stefan is also impossible to fully compensate for, as the love of a particular person is irreplaceable. Given the nature of their relationship, Elena has a special responsibility to Stefan (that third parties do not have) to care for his happiness—which is the source of her moral debt and her guilt—but this responsibility is overridden by her right to put her long-term well-being first.

Many might vehemently deny the suggestion that Sophie has a moral debt. It falls instead on the Nazis who are to blame for the child's death. However, it is possible that both Sophie and the Nazis have moral debts, though of a very different nature. The Nazis are indeed responsible for the child's death. Focus, however, on the meaning conveyed by Sophie's choice to her younger child: "You will be sacrificed to save your sibling". Sophie is the agent behind the decision that one child will die rather than the other. To repair the rupture created by her decision, she owes an acknowledgment to her younger child that she does not deserve to be sacrificed in this horrific way. Guilt is the natural human expression of this moral debt.

Sophie is thus doubly tragic: not only does Sophie lose one of her children, but she is also *forced* into the burden of guilt by the Nazi prison guards. While we can assume that Agamemnon has at least voluntarily taken up the heavy responsibilities of military command, Sophie is coerced into performing an intentional action—choosing a child to be killed—that fittingly stirs up guilt.

Sophie reveals that guilt can be fitting but horribly unjust. Others may impose all sorts of negative emotions on us, such as sorrow, disgust, anger, fear, shame—and it appears that guilt is no exception. The injustice of Sophie's guilt gives her reasons to let go of it despite its fittingness.

It is worth noting that not all instances of justified harm create a moral debt. A doctor performing a painful surgery does not disrupt her relationship with the patient, as the pain is suffered for the sake of the patient's own well-being rather than someone else's. Moreover, a gay man's feelings of guilt for coming out would be unfitting despite his homophobic parents' disappointment and distress. He has brought about a disruption of the moral relationship with his parents but has no moral debt to them. He has not done anything morally wrong, and any emotional harm that he has caused his parents is due to their own unjustified prejudices.

4.6 Guilt and desert

I have shown how the Moral Debt Account can explain fitting guilt without blameworthiness, as well as many of the conflicting intuitions surrounding it. However, there is a worry that I have not yet addressed and that concerns many instances of guilt over innocent behaviour.

Andreas Carlsson argues that guilt is an emotion governed not by *fittingness* but by *desert*, spelled out in the following way: a deserved response is one (1) that is non-instrumentally good and (2) that others have a *pro tanto* reason to induce in an agent (Carlsson, 2017, p. 112). Guilt, however, is necessarily painful, and, intuitively, agents do not deserve to suffer for actions, attitudes, and omissions that are outside of their control (Carlsson, 2017, p. 91). Imagine that Damon's actions in ***Reckless driver*** were a result of someone secretly slipping him a drug that had made him extremely prone to taking risks. He would then not have deserved to feel guilty; it would have been unjust if others were to make him experience emotional pain for his actions, since they were outside of his control. Carlsson's argument creates a problem for the Moral Debt Account. This is because it is beyond the control of many of the morally innocent agents that I have been discussing (like Tyler, the lorry driver, Sophie, and Agamemnon) to incur a moral debt. Agamemnon, for instance, finds himself in a dilemma: he will either have to sacrifice his daughter or fail in his duties as a commander.

There are two ways to respond to this challenge. The first one is to simply deny that guilt is exclusively governed by desert. It is fitting, and even non-instrumentally good (i.e., satisfying (1)), for innocent agents to feel guilty when they have a moral debt that disrupts a relationship, but others do not have a *pro tanto* reason to induce it in them (i.e., not satisfying (2)). On this picture, guilt is analogous to grief (Moller, 2007): both emotions can be non-instrumentally good but cannot be demanded by others. Guilt is fitting some of the time and perhaps deserved only when the agent is blameworthy, as blameworthiness licenses bringing about painful emotions. Others may induce guilt in Damon but not in Tyler, the lorry driver, and the rest. Once guilt is distinguished from blameworthiness, as it is on the Moral Debt Account, there is nothing that stands in the way of claiming that guilt is deserved only when the agent is blameworthy.

Yet, some may be dissatisfied with this view because of the intuition that guilt is at least sometimes deserved by morally innocent agents in virtue of their moral debts. We may judge, for instance, not only that it is non-instrumentally good for Agamemnon to emotionally acknowledge his horrific moral debt, but also that the debt is so great that we are entitled to induce guilt in him. It feels natural to exclaim, "How can you not feel guilty for killing your own daughter?".

There is a second way to respond to Carlsson's argument that accommodates this intuition. It relies on the idea that there can be various kinds of desert. This move should be familiar by this point in the monograph. Let us assume that desert, as Carlsson defines it, just is a normative relation that entails (1) the non-instrumental goodness of a response and (2) a *pro tanto* reason to bring it about. Then, any given response may be deserved or undeserved along several dimensions, as there are different kinds of non-instrumental goodness and different reasons to bring those about. There is no such thing as *the* desert. When it comes to guilt, we can distinguish between two different kinds of desert, each stemming from a different moral principle. The first one we know well by now:

> *FAIRNESS*: People should not be treated differently, in terms of burdens and benefits, for what is outside of their control.

The second one is:

MORAL DEBT: Agents deserve to experience a pained acknowledgment of their moral debts, incurred in virtue of a disrupted relationship, in the form of guilt.

While FAIRNESS is a broad principle that applies to many moral issues, *MORAL DEBT* is instead a distinct principle that is specific to guilt and its relation to moral debts. Just like, intuitively, grief is a non-instrumentally good response to a great loss that has nothing to do with the goodness of fairness, so guilt is a non-instrumentally good response to moral debts, and it can sometimes give reasons (not stemming from fairness) for others to bring it about.

Distinguishing between *FAIRNESS* and *MORAL DEBT* makes it possible to have more nuanced judgments about guilt. In particular, Agamemnon's guilt can be unfair (i.e., not deserved according to *FAIRNESS*) and yet an appropriate acknowledgment of his moral debt (i.e., deserved according to *MORAL DEBT*). If this is correct, then the pain of guilt will, tragically, not always be fair. The right attitude towards Agamemnon would be one of ambivalence: we would have reasons counting both for and against making him feel guilt for sacrificing his daughter, and there would be a moral cost in either case.

4.7 Conclusion

I have proposed that guilt targets not blameworthiness, but rather, more broadly, moral debts, and I have explored three kinds of situations in which one may incur a moral debt despite not being blameworthy. An innocent agent may owe others to restore equity, to commit to making amends when in a state of uncertainty, or to acknowledge and compensate for an all-things-considered justified harm or wrong. Many cases found in the literature on guilt without blameworthiness fall under one of these three kinds. The rest can perhaps be explained by other kinds of moral debt. For example, we may have moral debts stemming from our close ties and group memberships; this can account for the guilt felt for the actions of one's Nazi grandfather.

Though the Moral Debt Account expands the sphere of fitting guilt to certain blameless agents, there are numerous reasons that may count against their actually experiencing guilt. As I have shown, guilt can be pointless even when fitting and unfair even when deserved (along one dimension). In addition, guilt can be a traumatically painful emotion; it can also be easily abused, as it is a form of submission to the power of others. These and other considerations

should make us wary when we make judgments about guilt in particular circumstances.

In virtue of seeing blameworthiness as separate from fitting guilt, the Moral Debt Account has implications for several debates in which fitting guilt figures as a central element. I will briefly mention two. First, proponents of the existence of moral dilemmas—situations in which every course of action results in the agent's being blameworthy—rely heavily on emotions like guilt to make their case (Szigeti, 2015). According to some, the conclusion that Agamemnon and Sophie will be blameworthy whatever they do follows from the intuition that they will feel guilty whatever they do (e.g., Greenspan, 1983b). From the point of view of the Moral Debt Account, this argument relies on an equivocation. The emotion of guilt does not imply culpability but rather a moral debt in virtue of a disrupted moral relationship.

Second, the existence of outcome moral luck is also partly explained in terms of negative self-directed emotions like guilt (Lang, 2021; Nagel, 1979, p. 16; see also Wolf, 2001). An unlucky drunk driver who hits a pedestrian on his way home would feel much more guilt than a drunk driver who gets home safe because, luckily, he encounters no one. Therefore, defenders of outcome moral luck claim, the lucky driver is less blameworthy than the unlucky one. The Moral Debt Account allows instead for a view on which the two drivers are equally blameworthy though they differ in their moral debts. The lucky driver, in contrast to the unlucky one, does not stand in a disrupted relationship with an injured pedestrian, and hence there is less need for him to repair and signal submission through his guilt. Outcome moral luck, moral dilemmas, and other puzzles in moral philosophy can be approached with more subtlety if guilt is recognized as much more of a social emotion that serves as a signal when relationships are disrupted than as the punishment meted out in the soul's internal courtroom.

Chapter 5. Regret and indirect moral luck

5.1 Introduction

In the first three chapters, I applied the concept of ambivalence to argue that there are moral costs whatever we do in contexts where we need to respond to others' actions. On the one hand, it is (ultimately) unfair to respond in unwelcome ways to agents whose actions are (ultimately) outside of their control. On the other hand, failure to do so would result in giving up other valuable concerns, which stem from retributive, instrumental, and emotional considerations.

The concept of ambivalence, however, can shed light on another set of questions about responsibility. How should we hold agents responsible who make choices that merit ambivalence? I will consider two issues related to this question. In Chapter 6, I will focus on agents who are not morally blameworthy but nevertheless bear responsibility for choosing supererogatory demands over their personal life or vice versa. But first, in this chapter, I will turn to agents who are morally blameworthy and yet justified in a broader sense—an issue brought up by Bernard Williams (1981b) in "Moral luck".

Williams's (1981b) and Tomas Nagel's (1979) papers for the symposium on moral luck in 1976 founded the contemporary subdiscipline in philosophy devoted to understanding how much, if at all, moral evaluations are affected by forces outside of an agent's control. While both authors argue that morality is indeed disconcertingly subject to luck, they set out to explore the issue in fundamentally different ways. Nagel maintains that moral evaluations are *directly* affected by factors beyond an agent's control: her constitution, her circumstances, the consequences of her actions, and the causes of her actions (§1.2.1).

Williams, however, while not disagreeing with Nagel, pursues a different strategy. He does not target moral evaluations directly. Rather, he has puzzled philosophers for decades by claiming that *rational justification* is subject to luck, and that this entails that moral justification is, though indirectly, subject to luck too. In Williams's central example, whether Gauguin will turn out to be rationally justified for leaving his family to pursue an artistic career in Tahiti depends on whether he turns out to be a great artist or not, which, at least from Gauguin's epistemic point of view, is a matter of luck.

Philosophers interpreting "Moral luck" are sceptical of Williams's indirect moral luck strategy (e.g., Enoch & Marmor, 2007; Lang, 2018; R. J. Wallace, 2013). In this chapter, I will point out overlooked elements of Williams's text—which add to its plausibility—and suggest some amendments in order to make a stronger case for indirect moral luck. I will, in the process, bring to the surface questions about responsibility that Williams thought were important but have not received enough attention in the literature, such as: Can we make sense—and not just conceptually, but in terms of real-life experience—of a person's being morally unjustified and yet justified in a broader sense, as a rational agent? Can there be good reasons to choose to be blameworthy?

I will first set the stage by outlining Williams's aims and opponents in "Moral luck" (§5.2). I will then offer an interpretation of the case of Gauguin that brings out Williams's overlooked criticisms of John Rawls's (1971) notion of a rational life plan and of a philosophically dominant view of regret (§5.3). Finally, I will turn my attention to the issue of indirect moral luck (§5.4).

5.2 Regret and rational plans of life

Williams states that one of his main aims in "Moral luck" is to argue that *rational justification* is subject to luck (Williams, 1981b, p. 24). It will be helpful, however, to approach Williams's paper by avoiding, at first, the concept of rational justification, and distinguish instead between three interrelated aims that Williams pursues in the text: (1) to criticize a philosophically dominant conception of rational self-reproach in the form of regret, (2) to criticize John Rawls's (1971) notion of a *rational life plan,* and (3) to demonstrate that moral justification is indirectly subject to luck. The argument for the first two depends on the phenomenon of transformation of an agent's evaluative outlook, for which Gauguin serves as the central example.

Before we turn to Gauguin, however, we must first look at regret. As I mentioned in §4.2, Williams relies on the concept of "agent-regret" to serve his argumentative purposes in "Moral luck", but there are serious doubts about whether agent-regret, at least as characterized by Williams, is a distinct emotion (Jacobson, 2013). The constitutive thought of agent-regret is "I wish that I had done otherwise" (Williams, 1981b, p. 27), but this is too broad, as it can include feelings of guilt (e.g., in the case of the lorry driver (§4.3).

However, this hurdle need not obstruct Williams's arguments. Some minor alterations of his account of regret can set us on the right track. Agent-regret is too broad, but it can be narrowed down. There is indeed a distinct natural emotion, quite close to agent-regret, that targets one's own rational errors. We may call this *agential regret* (Jacobson, 2013, p. 104; Shoemaker, 2015, p. 67). Agential regret—regret over one's own exercise of agency—is commonly thought of as the "cousin" of guilt in the sphere of rational action. Agential regret is, essentially, emotional self-reproach about one's own rational decisions. It is a painful emotion that motivates dwelling on one's own rational errors and intending to act differently in the future. For instance, my rash decision to invest in cryptocurrencies that leads to my losing a lot of money may fittingly evoke agential regret.

Since agential regret targets only the agent's own errors, when we say that we regret events or other people's actions (e.g., "I regret that WWII happened"), we are actually experiencing sadness, dismay, or anger rather than agential regret (Jacobson, 2013, pp. 100-101; cf. R. J. Wallace, 2013, p. 33). Regret over one's *own* errors is a distinct emotion and has special importance in one's life because it is a form of self-reproach. Agential regret involves the sting of the thought that *I* was at fault.

The main line of Williams's argument can be pursued by substituting agent-regret with agential regret. With this amendment, we can return to Williams's goals in "Moral luck". One of his tasks is to argue against a view according to which *it is wholly within our control to avoid (fittingly) experiencing agential regret*. Influential philosophers like Rawls claim that an agent who deliberates impeccably at the time of action, basing herself on what is known to her then—we may call this *ex ante deliberative rationality*—will not experience agential regret over her decision even if things happen to turn out badly (Rawls, 1971, p. 422). Imagine that, after flawlessly deliberating about it, I decide to pursue a career in engineering. If the state of the economy unpredictably changes, and I

end up unemployed, I may be sad about the outcome but will not have grounds for self-reproach. As long as I consistently perform well as an *ex ante* rational deliberator, I will be free of (fitting) agential regret.

Though at first glance plausible, this view of regret, Williams claims, rests on a false presupposition:

> The presupposition can be put like this: as rational agents, we seek to be rational; to the extent that we are rational, we are concerned with our agency and its results to the extent that they can be shaped by our rational thought; to the extent that results of our agency could not be affected by greater rationality, we should regard them as like the results of someone else's agency or like a natural event. (Williams, 1995c, p. 245)

In short: the dominant view of agential regret rests on the assumption that it is fitting to reproach oneself only if one could have done better at the time of deliberation. Everything that is not an *ex ante* rational failure should be viewed "as like the results of someone else's agency or like a natural event". Williams thinks that this is false: an agent, like Gauguin, may fittingly experience agential regret even when he has deliberated flawlessly at the time of choice.

Williams argues that the misguided view of regret is a result of the "moralization of psychology" (Queloz, 2021a): the problematic philosophical tendency to misrepresent human psychology in order to fit with moral ideals. The belief that it is within our control to be free of agential regret has a long history in philosophy. It is found already in ancient Stoicism. At the core of Stoic philosophy is the idea that "the virtuous person is fully rational", and "since the virtuous person gets it right, she will not be troubled by regrets. . . this may leave room for some kinds of regret, but not regret for having done the wrong thing" (Annas, 1993, p. 409). In other words, the Stoic sage is free of agential regret. This dictum is part of the larger "extreme" Stoic philosophical system, according to which one's happiness (eudaimonia) is entirely within one's control (Williams, 1981b, pp. 20-21). Luck does not threaten the happiness of the rational person.

For Williams, traces of this Stoic resistance to luck are found in the contemporary Western moral philosophical outlook—manifested in theories like Kantianism and utilitarianism—that he terms "the morality system" (Williams, 2011, Chapter 10). While the morality system admits that much of human happiness is tragically at the mercy of luck, it nevertheless finds a

sanctuary, a shelter from luck, in morality (Queloz, 2022; Russell, 2017). A person's moral worth, from the perspective of the morality system, is of great importance and impervious to luck, as whether she will be (appropriately) blamed or praised is based entirely on actions that are under her control.

Commentators of "Moral luck" point out that it is an attack of the morality system's idea that moral evaluations are sheltered from luck (Lang, 2018, p. 131; R. J. Wallace, 2013, p. 136; Williams, 1981b, pp. 21-22), but, it bears emphasizing, Williams has multiple aims in the text. He believes that the morality system's resistance to luck leads not only to a flawed theory of morality, but also to a flawed theory of rationality (Williams, 1995c, p. 246). One aspect of the latter is the misguided account of agential regret. A closely related wrong idea about rationality, for Williams, is Rawls's conception of a rational life plan.

In *A Theory of Justice*, Rawls (1971) presents the concept of a rational life plan, which is used to set a criterion for the good of each individual person. What is good for a person is fixed by what is rational for her to want, and hence what is rational for her to plan for—an idea found in Aristotle, Thomas Aquinas, Kant, and Sidgwick, among others (Rawls, 1971, p. 400, fn. 2). Rawls's concept of a rational life plan is intimately bound up with his account of agential regret. This is because to live in accordance with one's rational plan of life, an agent must abide by "the guiding principle that a rational individual is always to act so that he need never blame himself no matter how things finally transpire" (Rawls, 1971, p. 422). If a person follows a plan for her life constrained by the principles of *ex ante* deliberative rationality, she will be free of agential regret, i.e., regret over her own conduct. If, in addition, external circumstances are favorable, the person following her rational plan of life will be happy (Rawls, 1971, p. 409).

Williams argues not only that following a rational plan of life cannot secure freedom from regret over one's conduct, but also that the concept of a rational life plan cannot fulfill its function, i.e., fix a person's good. Both conclusions follow from the observation that people go through evaluative transformations throughout their lives, which is illustrated by the case of Gauguin.

5.3 Gauguin

In this section, I will offer an interpretation of Williams's Gauguin case that can make sense of the first two of his aims in "Moral luck", namely (1) to criticize a

dominant philosophical conception of regret, and (2) to criticize the closely related Rawlsian idea of a rational life plan.

Many share the intuitions about agential regret and the rational plan of life that serve as the basis of Rawls's theory. Williams, however, is distrustful of such intuitions and proposes instead a different philosophical method: to reflect on some special situations "not in terms of substantive moral opinions or 'intuitions' but in terms of the experience of those kinds of situation" (Williams, 1981b, p. 22). Williams's description of two cases—involving Gauguin and Anna Karenina—are an application of this method. I will focus on the case of Gauguin since it is more extensively discussed by both Williams's and his commentators.

R. J. Wallace (2013) offers a detailed construal of the case of Gauguin by making use of a broad concept of regret. For Wallace (2013), regret is "an active wish that things had been otherwise in the relevant respect" (p. 78). Notice that this kind of regret is not equivalent to the distinct natural emotion of agential regret because it is not exclusively focused on the quality of a person's own agency. For Wallace, I may regret the weather because I have an active wish that it had been different; or I may regret my choice because it led to an unfortunate outcome and not because *I* was somehow at fault. Regret, as defined by Wallace, appears to pick out a pained motivational state rather than a distinct emotion. An active wish that things had been otherwise can be accompanied by a variety of emotions, such as sadness, lamentation, anger, etc.

With this notion of regret at hand, Wallace proposes the following timeline of the Gauguin case. Gauguin decides to leave his family to pursue an artistic career in Tahiti. At the time of choice, he does not yet know whether he will turn out to be a great artist or not. The outcome, from his perspective, is at least partly a matter of luck. According to Wallace, if Gauguin succeeds as an artist, he will have a new valuable object in his life—artistic success. Because he is happy that he has achieved this success, he would not have an active wish to have chosen otherwise; he would be rationally unable to regret the decision to leave his family. Given how much importance Gauguin attaches to his success as an artist, he would affirm the necessary steps to realize it, which include leaving his family. In contrast, the unsuccessful Gauguin would have no valuable object to ground the affirmation of his decision. He would be left simply regretting leaving his family, a decision that has brought him nothing of value (Wallace, 2013, pp. 134-140).

Wallace, rightly, points out that this kind of regret over one's decision has little to do with judging whether it was rationally justified. A person who gambles all his retirement savings at risky odds and wins is not rationally justified just because he does not regret his gamble (Wallace, 2013, pp. 110-111). If this was Williams's argument for the claim that rational justification is subject to luck, it would be clearly unsuccessful. Rational justification, Wallace claims, is established by whether the agent acts in accordance with *ex ante* principles of rationality regardless of the outcome. Therefore, an agent like Gauguin or the gambler should have mixed feelings about his choice: he should affirm his decision because it brought into existence a new object of value in his life (e.g., artistic success, money), but at the same time he should regret the decision because it was rationally unjustified (Wallace, 2013, pp. 169-ff). It was a rationally bad decision that fortunately turned out well.

Williams's argument is indeed demonstrably weak on Wallace's interpretation. However, here is an alternative. Successful Gauguin does not merely get a new object of value in his life (artistic success) but now transforms in such a way that he *comes to value the objects in his life differently.* He goes through a transformation in his evaluative outlook: "The outcome [of Gauguin's choice] has to be substantial in a special way—in a way which importantly *conditions the agent's sense of what is significant in his life*, and hence his standpoint of retrospective assessment" (Williams, 1981, p. 36, my emphasis).[1] Successful Gauguin would have become someone who can confidently judge that his life's project is valuable. Now he is someone who thinks that he was right to choose to depart for Tahiti because his art was worth the tragic sacrifice. Conversely, the unsuccessful Gauguin transforms into a person for whom art is a meaningless pursuit. An unsuccessful Gauguin does not just regret what happened to follow from his decision, as if it was just any event capable of producing good or bad results. He regrets *his choosing,* the *quality* of this decision, because from his new evaluative standpoint he sees it as profoundly wrong: he set a meaningless art career against the happiness of his family.[2] If he

[1] This evaluative transformation appears to be close, if not identical to, Laurie Ann Paul's notion of personally transformative experience (Paul, 2014, p. 16).

[2] There could be resistance to the idea that it is a matter of luck whether Gauguin should feel agential regret over his decision because of the nature of the example. Many would judge that the moral cost of leaving one's family outweighs Gauguin's artistic success, and he should therefore regret his decision regardless of whether he turns out to consider his artistic career meaningful or not. I have sympathies with this view, and Williams acknowledges that intuitions here may vary

was confronted with a similar choice in the future, successful Gauguin would again choose his artistic career over a moral value like the happiness of his family. Unsuccessful Gauguin, in contrast, would never choose an artistic project again over anything because he has come to assign it no value at all.

Williams uses the phenomenon of evaluative transformation to make two interrelated points. The first is that it would be natural for an unsuccessful Gauguin to feel a self-directed form of regret over his own decision. An unsuccessful Gauguin would be too inhumanly detached from his own agency if he were to judge that his decision simply had a bad outcome; he would rather think that, from his new perspective, he made a stupid decision.

Perhaps the clash of the moral and non-moral in the Gauguin case draws attention away from the phenomenon of evaluative transformation. Here is a simpler example that can bring this point into relief. Imagine that in her youth, Roberta values her active pursuit of academic philosophy, but, later in life, she becomes a person who finds philosophy quite boring and wishes that she had spent her twenties travelling the world instead. Neither her younger nor her older self has made any mistake in *ex ante* deliberative rationality; she just happens to evaluate things differently at different times in her life. Nevertheless, she experiences deep agential regret about her pursuit of philosophy, feeling that she has wasted years in what she now sees as a meaningless project.

We can put the general point in the following way. A decision has both *form* and *content.* The form is made up of the *ex ante* rules of rationality, such as considerations about the "consistency of [the agent's] thoughts, the rational assessment of probabilities, and the optimal ordering of actions in time" (Williams, 1981b, p. 31; see also Richards, 1971). The content is made up of the objects of value that the rules are applied on; for Gauguin, the content of his decision is his artistic project and the happiness of his family. He assigns them a certain value at the time of choice, but his valuation changes as he goes through an evaluative transformation, and so, from a later evaluative perspective, he may see the earlier valuation as one that he can no longer endorse. Williams's example shows that agential regret may attach itself not only to the form of the decision, but also the content. Given the motivational profile of agential regret I outlined in the previous section (§5.2), this is not a surprising

(Williams, 1981b, p. 37), but it does not undermine the general point that agential regret over one's choices can be subject to luck.

result. Agential regret facilitates learning from one's mistakes, and there is no reason to suppose that it must only target mistakes in applying the formal principles of rationality. An agent can also fittingly experience agential regret over a former evaluative outlook, which would help her affirm and act in accordance with her new values as she changes throughout her life. Agential regrets could even be seen as partly *constitutive* of transformations in evaluative outlook: to abandon one's former system of values just is, in part, to regret having abided by it.

Wallace (2013, p. 169) and Lang (2018, p. 142) observe that the successful Gauguin should feel ambivalent about his choice because he affirms his success and yet regrets the moral sacrifice necessary for it. My analysis of the case reveals that there is also a source of mixed feelings for Gauguin, centered solely on his own agency. If he is sufficiently sensitive, an unsuccessful Gauguin would feel split not only about the tragic circumstances of his choice, but also about how he should judge *himself.* After all, assuming that he could not have done better at the time of his choice, he will stand by his *ex ante* deliberative process but regret his former evaluative outlook. On the one hand, he judges that he would apply the same rules of rationality in similarly risky circumstances. On the other hand, he now thinks that he should never again choose to pursue an art career over other objects of value, like his family's happiness.

To sum up, the first of Williams's points is that we do not always have control over whether our decisions turn out to be the kind of mistake that merits agential regret. Impeccable choices at an earlier time may *become* mistakes, at least regarding their content, for the agent at a later time. The Stoic sage is revealed to be fiction.

Williams's second point is that the concept of a rational life plan cannot serve its purpose in Rawls's theory. What is good for a person cannot be derived from a rational plan of her life because there is no single such plan that is relevant. Gauguin's case illustrates that our perspectives on what is valuable, and hence good for us, change throughout our lives.[3] The point is not merely the epistemic one that we cannot know in advance what would be good for us from a future perspective. It is in the nature of the example that Gauguin cannot predict how he will transform, but perhaps some evaluative transformations are foreseeable. The problem, rather, is that there is no clear way of deciding which standpoint

[3] Larmore (1999) develops a similar argument against Rawls.

of assessment (the earlier or the later one) to prioritize. What would the rational plan for Roberta's life be? Is it good for her to do philosophy, as her younger self prefers, or go travelling instead, which is what her adult self would find more worthwhile? There is no such thing as *the* rational life plan for a person's life:

> [W]hat one does and the sort of life one leads condition one's later desires and judgments. So there is no set of preferences both fixed and relevant, relative to which the various fillings of my life-space can be compared. . . The perspective of deliberative choice on one's life is constitutively *from here*. Correspondingly the perspective of assessment with greater knowledge is necessarily *from there,* and not only can I not guarantee how factually it will then be, but I cannot ultimately guarantee from what standpoint of assessment my major and most fundamental regrets will be. (Williams, 1981, pp. 34, emphasis in original)

Why should my interpretation be favored Wallace's? First, it explains why Williams insists that only a special kind of luck is relevant for Gauguin's judgment over the rationality of his decision and accordingly for his agent-centered form of regret. It is luck *intrinsic* to the project—in this case, whether Gauguin turns out to be a successful artist—that can justify or unjustify his decision (Williams, 1981, pp. 25-26). If Gauguin is unsuccessful because he falls victim to *extrinsic* luck (e.g., his ship capsizes on the way to Tahiti), he would never learn if his choice was justified or not. On Wallace's interpretation, the difference between intrinsic and extrinsic luck cannot be important, since what determines the agent's regrets is only whether there is a new object of value that causally follows from the choice. Neither the Gauguin whose ship capsizes nor the Gauguin who turns out to be a bad painter would become successful, and so, according to Wallace, both would regret their choices in a similar way. Wallace's account cannot make the difference between them. On the interpretation I am offering, intrinsic luck, unlike extrinsic luck, will lead to an evaluative transformation.

Another reason to favor my interpretation is textual evidence. Williams clearly states that Gauguin undergoes a transformation in evaluative outlook. Recall that, for Williams, the outcome "conditions the agent's sense of what is significant in his life, and hence his standpoint of retrospective assessment"

(Williams, 1981b, p. 36).[4] Wallace's interpretation is also at odds with what Williams explicitly says about regret. Williams acknowledges that *ex ante* deliberative rationality is standardly thought to determine rational justification in philosophy, but he brings into question whether it is the only way, or only significant way, in which we evaluate the rationality of our own decisions (Williams, 1981b, p. 35). He believes that we may regret our own choices by focusing neither on *ex ante* deliberation nor on the unfortunate outcomes that a choice may bring (Williams, 1995c, p. 245). Unsuccessful Gauguin is experiencing agent-regret, which is constituted by the *agent-centered* judgment "I wish that *I* had done otherwise", and not "I wish that things had been otherwise".

Finally, and connected to the previous point, consider that on Wallace's account, Williams's criticism of Rawls would be too evidently unsuccessful. Rawls is explicit that the form of regret precluded by following one's rational life plan is not regret about how things turn out but about the quality of one's own decision-making (Rawls, 1971, p. 422; Williams, 1981, p. 34; 1995, p. 245).

Interpretation aside, let us return to the issue of *rational justification* that I set aside in the beginning. I will not take a stand on whether Williams is correct that Gauguin's evaluative transformation affects how he views the rational justification of his decision. This depends on how broad we want the concept of rational justification to be. Maybe it should refer only to evaluations of *ex ante* deliberation and exclude subsequent revisions due to changes in the agent's value system—or maybe not. Williams's two criticisms—about agential regret and about the concept of a rational life plan—hold regardless of one's thoughts on rational justification.

5.4 Indirect moral luck

Back to the central question of "Moral luck": What does Gauguin's case have to do with *moral* luck? Here, Williams's argument becomes quite confusing, for

[4] An interpretation by Agata Lukomska (forthcoming) uses Williams's (1981c) concept of ground project to illuminate the case of Gauguin. For Lukomska, Gauguin's artistic career is a ground project, i.e., it is what makes Gauguin want to keep on living, and so his regret if he fails will be of an even deeper kind. Lukomska's account, however, still views Gauguin's decision in the light of whether the outcome brings an object of value (his ground project) in his life or not. It does not address the issue of evaluative transformation.

several reasons. Suddenly, evaluative transformation stops playing a role in the text. Williams claims that what matters to "the moral spectator", i.e., for moral justification, is whether Gauguin succeeds in bringing "a good for the world" (Williams, 1981b, p. 37) through his paintings. A moral spectator "has to consider the fact that he has reason to be glad that Gauguin succeeded, and hence that he tried" (Williams, 1981b, p. 37). Gauguin would be redeemed if leaving his family turns out to result in some wonderful art. The relevant feature of the outcome, Williams claims, is whether a new object of value appears, which has little to do with whether he undergoes an evaluative transformation.

As commentators have pointed out, this is a dubious argumentative strategy (Lang, 2018, pp. 133-134; R. J. Wallace, 2013, p. 172). An easy response for the denier of moral luck is that we may be grateful *that* Gauguin's actions brought about a good thing, but not *to* him. It is consistent to blame someone for a moral failure and yet be happy that this failure turned out to produce something of value. An accurate moral assessment of Gauguin would consider only what was in his control at the time of choice.

But this is not the only issue that makes Williams's thoughts on moral luck confusing. Besides unexpectedly dropping the idea of evaluative transformation, he also offers two possible conclusions about Gauguin's moral justification. One is that Gauguin's choice to leave his family becomes *morally justified* by his success as an artist (despite, even more confusingly, its being appropriate for his family to blame him). On this reading, Gauguin's case is not one in which there is merely "gratitude that morality does not always prevail—that moral values have been treated as one value among others, not as unquestionably supreme" (Williams, 1981b, p. 37). The successful Gauguin's art *morally* redeems his decision.

On the other hand, Williams supposes the exact opposite a few passages later:

> [P]erhaps we should, all the same, accept that conclusion. Their moral luck, we should then say, does not lie in acquiring a moral justification. It lies rather in the relation of their life, and of their justification or lack of it, to morality. (Williams, 1981b, p. 39)

The second possible conclusion, according to this passage, is that although Gauguin is *not morally justified*, he is justified in a broader, all-things-considered sense; he is justified on "the ultimate and most important level" (Williams, 1981b, p. 22; see also Williams, 1995c, p. 245). On this higher level of agential

assessment, *all* values, not merely moral ones, are taken into account. Gauguin is morally blameworthy for leaving his family, but he would be justified in affirming his choice overall—and, Williams supposes, so would others who are able to take on his perspective. This seems to be the alternative that Williams favors, as he returns to it in "Moral luck: a postscript" (Williams, 1995c, pp. 244-245).

Both alternatives are sketchy and raise more questions than they answer. Rather than stop here, however, I will show that, though he does not explicitly make use of them, Williams's text provides the elements to make a stronger case for indirect moral luck. Instead of focusing on whether Gauguin's decision results in a good for the world that one should be grateful for, we should rather consider how Gauguin's evaluative transformation affects the moral evaluation of his decision. To get a grip on the many moving parts, I will proceed by asking two key questions in turn.

First: Can we make sense of a person's being morally unjustified and yet justified in a broader sense, as a rational agent? Let us put to the side the Gauguin case and consider a different example:

> ***Business venture.*** Donna has promised Tanya to open a hotel together on a tropical island. However, Donna suddenly discovers that she would feel profoundly homesick and unhappy if she embarks on a business venture in a far-off place, and so she breaks her promise to Tanya. The effort and resources that Tanya has put into the business go to waste and she feels betrayed by Donna.

Donna faces a special kind of *ambivalence*, where she must choose between keeping a promise and her happiness; she chooses the latter. What would be the appropriate response to Donna? Let us suppose that Donna is somehow correct in her decision. Consider the two alternatives proposed by Williams: Donna is either (1) morally justified or (2) not morally justified but instead justified in an all-things-considered sense. If Donna is morally justified, then Tanya (and others) would not be entitled to hold her accountable. It would be inappropriate, for example, to resent Donna or to subject her to the unwelcome effects of blame. This does not appear plausible in ***Business venture***. Tanya appears entitled to hold Donna morally responsible for her betrayal.

The second alternative is that Donna is morally blameworthy for breaking her promise—it would be appropriate to resent her, ask for an apology, etc.—but,

at the same time, she is rationally justified in doing so. This means that given all the values at stake, it is rational for Donna to prefer her own happiness at the expense of committing a relatively minor moral wrong.[5] It is better for Donna to be blameworthy and happy than to be not blameworthy and unhappy. Furthermore, let us imagine that others, including Tanya, agree that Donna is, all things considered, correct to prefer her happiness over keeping the promise. Tanya understands that if she were in similar circumstances, she would do the same. Put differently, Donna has a valid, intersubjectively recognized reason to *choose to be blameworthy*. Tanya, therefore, simultaneously blames Donna (and demands an apology, some form of compensation, etc.) and affirms Donna's choice.

Is this picture psychologically and ethically plausible? My goal here is to present the question rather than answer it conclusively. The idea that one can be rationally justified but morally unjustified is taboo in many areas of moral philosophy, where moral justification is viewed as supreme; if one is justified as a rational agent, then one must *ipso facto* be morally justified too. I will not argue against such views, but it is nevertheless worth asking how ethical and psychological life would be like if they were wrong.

Let us assume that the answer to the first question is "yes" (we can make sense of a person's being blameworthy and yet rationally justified). We may then ask: Can an agent's evaluative transformation affect how she views the rationality of her past choices, and thereby, indirectly, how she views the importance of their moral worth?

To stick with ***Business venture***, imagine that years later, Donna transforms into a person who gets over her homesickness; she is now awfully tired of her hometown and craves going abroad. Accordingly, Donna now feels agential regret about her decision to back out of her commitment to Tanya. Before, she weighed the value of the alternatives correctly, but from her new perspective, she sees the weighing as wrong. She was blameworthy, but she could affirm her choice; now, she is still blameworthy, but cannot affirm her choice anymore.[6]

---

[5] "Rational" is not equivalent to "egotistic". Donna, and others evaluating her choice, care about moral values but treat them as only one dimension of value.

[6] Donna can still affirm her *ex ante* reasoning (the form of the decision), but let us leave this to the side and focus on her affirmation or rejection of the valuation of the objects at play (the content of the decision).

This is a case of indirect moral luck. The agent's degree of blameworthiness for her choice does not change, but whether she would affirm her choice overall changes with time due to an evaluative transformation that is outside of her control.

Indirect moral luck appears to have the most importance for the agent's own relationship to herself, as it seems to have limited interpersonal effects. After all, Donna remains blameworthy even after her evaluative transformation. What about the emotional responses of others to the *rational* (and not moral) aspect of her choice? In the moral sphere, human emotions track both one's own and others' failures: we experience resentment towards others' wrongdoing and guilt towards our own. But when it comes to the rational quality of decisions, we appear to be more emotionally attuned to our own failures than to those of others. We experience intense agential regret when we make rational errors, but when others make such mistakes, our emotions appear to be less strong. We may feel, perhaps, at best, *disappointed* in another because of their rational error (Shoemaker, 2015, p. 66). Nevertheless, even if others do not have strong emotional responses to what Donna now sees as a mistake, they would presumably change their judgments about it. In the past, Tanya blamed Donna, but still saw her decision as the better option for Donna, all things considered. Now, Tanya still thinks that Donna is blameworthy for her choice to break her promise but acknowledges that Donna is now left with no grounds for overall affirming her choice.

5.5 Conclusion

Williams raises numerous issues in "Moral luck". This makes it a fascinating text, but one that is difficult to follow and evaluate. I distinguished between three interrelated aims that Williams has: a criticism of a dominant view of regret, a criticism of the notion of a rational life plan, and an argument for the existence of indirect moral luck. By spelling out the nature of Gauguin's evaluative transformation, I argued that Williams succeeds in his first two aims.

As to the existence of indirect moral luck, Williams unexpectedly offers a line of argument that is disconnected from the phenomenon of evaluative transformation, to which much of his text is devoted. I agreed with critics of Williams that gratitude for Gauguin's art should not affect how he and others view the *quality* of his decision. However, Williams sets the stage for another way to argue for indirect moral luck, though he does not take that path. Agents

like Gauguin can undergo evaluative transformations, which affects whether they would affirm, all things considered, their decisions *qua* decisions (rather than *qua* events necessary to bring about objects of value). Indirect moral luck occurs when an agent changes her evaluative outlook in such a way that she can no longer affirm her choice of another value over being blameworthy. Indirect moral luck, therefore, does not affect whether the agent is blameworthy; it affects instead the evaluative context, which dictates the importance of an agent's blameworthiness for her life. Much of my discussion on indirect moral luck is speculative; a lot of questions remain about the conflict between moral responsibility and other values in a person's life. This direction of thought is made possible by rejecting, like Williams does, the morality system's assumption that certain moral justification is the last court of appeal.

Chapter 6. Moral demandingness

6.1 Introduction

How much of my resources, time, effort, and well-being should I give up for a moral cause, such as helping the world's unfortunate people—the sick, the starving, the poor, the abused, the refugees? Half a century ago, this question was forcefully put on the philosophical agenda by Peter Singer (1972), who argued that, at the very least, "if it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it" (p. 231). If Singer is right, then almost all people in affluent countries are failing to live up to the demands of morality, which has sparked a debate over the more general question of whether morality can be overly demanding.

The debate has centered around drawing the line between the morally obligatory and the morally supererogatory in the right place. The difference between the two spheres is standardly spelled out in terms of responsibility: if you fail to do what is morally obligatory, you are blameworthy; in contrast, to perform a supererogatory act is admirable, but not performing it is neither blame- nor praiseworthy. According to some, morality would be overdemanding if it deemed big personal sacrifices obligatory rather than merely supererogatory.

In this chapter, building on insights about guilt from Chapter 4, I will put in doubt the assumption that there is a line that neatly divides the sphere of the morally obligatory from the supererogatory. I will suggest that the problem of moral demandingness has the shape of a particular kind of ambivalence (call it: *the moral/personal dilemma*). While it is sometimes permissible to choose one's personal life over a moral cause and vice versa, both choices may come with an ethical cost. One must either sacrifice one's autonomy and personal projects or must otherwise fail along a moral dimension, taking on the burden of guilt and disappointment for choosing oneself over others. Deciding in favor of one's

autonomy merits a complex responsibility response, which does not neatly fall into the general category of blameworthiness (nor even into the more nuanced categories of attributability or accountability (Macnamara, 2011; Shoemaker, 2015; Watson, 2004b)). I will also argue that, regarding responsibility, the moral/personal dilemma is substantially different from three similar phenomena, namely (1) the "classic" ethical dilemma where all available options breach a moral requirement, (2) cases of moral debt over all-things-considered justified moral costs (which I discussed in section §4.5), and (3) suberogatory acts (Driver, 1992).

I will first characterize the moral/personal dilemma as a conflict between moral demands on the one hand and the value of autonomy on the other (§6.2). I will then spell out the moral failure that ensues from choosing one's personal life over some kinds of supererogatory moral demands (§6.3). In §6.4, I will explain how the agent undergoing such moral failure merits certain responsibility responses (like guilt and reactive disappointment), but not others (like liability to demands and sanctions), and I will conclude in §6.5.

6.2 Autonomy and the limits of morality

The problem of moral demandingness is a general problem in moral philosophy that can be encountered in a variety of contexts. For example, one may feel the pressures of moral demandingness when one can help save a suicidal friend only at a great cost to oneself. I will, however, use the big global collective problems of the contemporary world (climate change, disease, famine, war, etc.) as the main examples in the text. It seems that many citizens of affluent countries face an ethical challenge that can be put as follows. Each day, people in the world die and suffer greatly, many through no fault of their own. Human activity is destroying nature and its living organisms on a grand scale. Yet, much of this destruction, death, and suffering is preventable, and you can help alleviate it. So why are you spending your free time playing video games instead?

There are several possible responses to this accusation. An all-too-common one is to claim that in fact it is not in one's power to help. This is implausible. True, many contemporary problems cannot be effectively tackled by throwing one's excess money at them (though some can be) and instead require changes in political and economic structures. But this observation does not absolve one from responsibility; it just means that helping must take forms that go beyond donation, like political action. Structural change can begin from individuals'

acts, and so morality is revealed as *even more demanding,* rather than less, since it requires not just your spare money, but also your focus, time, and effort.

Even when it is within one's power to help, perhaps morality's demands can be limited by shifting responsibility. One way to do this is through an appeal to fairness. The responsibility for global problems, one could argue, is collective and should be distributed equally among those who can help, and no one can be required to give more than one's fair share, even if in reality not everyone contributes (Murphy, 2000). In effect, this yields the moral obligation to give a relatively small fraction of one's resources that amounts to one's fair share of the burden and hope that everyone else does the same.

The argument has appeal, but it cannot serve to conclusively limit morality's overdemandingness. It appears that if a person truly cares about the values at stake (and not simply about dodging responsibility), she will often have to do more than her fair share. This is because in our morally flawed world very few collective action problems would be solved if it were morally required to do only one's fair share, since not everyone is aware of their obligations or is willing to comply. Some problems, such as the destruction of the environment and climate change, might require noticeable sacrifices from most and one cannot rely on others' readiness to give up that much. From what we can see in the world today, though unfair, one might be pressured to take on a relatively heavy burden because one must operate under the assumption that few agents would know that they need to contribute, what and how they need to contribute, and would then actually contribute. Furthermore, shifting responsibility to the people and institutions who must rightfully take it on is itself a burdensome act. One reason why morality is so demanding is that a lot of effort is required to figure out what must be done and to get others to do it too. "But I did my fair share!" seems like a weak excuse when explaining oneself to future generations who must deal with the consequences of our destructive consumption.

More importantly, however, the appeal to fairness only *contingently* secures an escape from moral demands. There could be times and places when preventing great suffering would require enormous effort and resources even if everyone does only their fair share. The relative effort and resources required from citizens of affluent nations to fight famine was perhaps greater fifty or a hundred years ago than it is now; and it might be even greater in the future. A fair distribution of burdens does not in principle limit moral demands to a comfortable degree. Again, even though I am using collective global problems as the main example

in the text, I am interested in moral overdemandingness as a general problem, which does not necessarily need to take on a collective form. Often it is as individuals that we are confronted with pressing moral demands.

Another attempt at limiting morality's demands is made by Bernard Williams (1981c) in his influential "Persons, Character, and Morality", where he puts forward what appears like a transcendental argument, though its details are somewhat obscure.

Williams introduces the concept of "categorical desires" (or "ground projects") to explain why we go on living: these are the desires that are not conditional on our existence, but rather settle the question of whether we have reason to exist at all (1981c, p. 11). While I want to be well nourished *if* I am to exist (a conditional desire), I exist *for the sake of,* say, developing as a musician (a categorical desire). My musical career gives me a reason to go on living. Williams then goes on to argue that

> [t]here can come a point at which it is quite unreasonable for a man to give up, in the name of the impartial good ordering of the world of moral agents, something which is a condition of his having any interest in being around in that world at all. (Williams, 1981c, p. 14)

But what makes it "quite unreasonable" for an agent to give up her ground project in such a case? On one interpretation, Williams is putting forward a transcendental argument that is supported by his theory of internal reasons (T. Chappell, 2007). Williams's Internal Reasons Thesis states that reasons for action can only be derived from a person's subjective motivational set *S*, which includes her desires, commitments, projects, etc. (Williams, 1981a; 1995b; see also §2.3). There must be a sound deliberative route from an agent's *S* to a particular action for there to be a reason for that agent to perform it. And if this is true, it would be self-defeating (and hence irrational) for an agent to prioritize moral demands over the ground projects that give her life meaning, since these projects are what keep her interested in acting in the first place (T. Chappell, 2007, p. 258). I cannot rationally give up my musical ambitions for the sake of a moral cause if my musical ambitions are what I go on living for.

This transcendental argument, however, runs into an obvious problem: we sometimes make moral demands on agents even when complying with those demands is irrational from the agents' perspectives. We would demand from a child-molester to drop his "ground project" even if that would be the only thing

that gives his life meaning and blame him if he refuses to do so. Such demands are not appeals to the agent's rationality but are rather acceptable forms of *coercion.* Williams himself acknowledges that we blame immoral individuals whom we cannot reach through appeals to act rationally by their own lights (the "hard cases"), despite this kind of blame being a mere "rejection" of their reasons (Williams, 1995b, p. 44).

Susan Wolf (1997) suggests a way to overcome this hurdle. According to her, there is an objective fact about which projects are meaningful:

> . . .even if we accept Williams' claim that morality cannot reasonably be expected to trump in cases where it conflicts with meaning-providing activities, this would not imply that, so to speak, anything goes. Were a child-molester to claim that his life would lose all meaning if he could not molest children, it would be in order to reply that if that were really true then his life would be meaningless anyway. Child-molesting, since it is lacking in value, is not the sort of thing that can give meaning to one's life. (Wolf, 1997, p. 306)

On Wolf's account, then, the transcendental argument yields that it would be irrational for a person to give up her ground projects for the sake of morality, the key detail being that "rational" should be understood as "objectively rational", rather than "rational for the agent".

However, once we step out of an agent's own perspective on her actions and into the domain of objective rationality, it could be objected that Wolf is begging the question against defenders of a more demanding morality. Once coercion is allowed—and it must be if child-molesters are to be kept in line—then one could argue that it is objectively rational for a person to give up whatever she subjectively sees as a meaningful project and instead adopt a moral project, as the latter is in fact the truly meaningful one. Defenders of a more demanding morality could argue that saving whales from extinction or helping alleviate world hunger is objectively the best and most meaningful thing a person can do with her life, and hence the most objectively rational.

Perhaps we do not need to settle what is most objectively rational or meaningful for a particular agent to do because Williams's transcendental argument should be understood on a more general level, as directed against a moral theory that issues *universal demands*. We should ask: What would happen if it was a *universal law* that, in certain conditions, one ought to give up all her meaningful

projects for moral causes? There would be nothing left but people living for each other's sake with no one doing the actual (meaningful) living. Since moral goals can be achieved only if there are some people who find their lives meaningful for non-moral reasons, and hence want to keep on living, morality would be self-defeating if it were ultimately and universally overriding. It makes sense to save a person's life only if she finds her life meaningful, and she must find it meaningful for a reason other than saving other people's lives. A moral theory can be overdemanding for some, such as the child-molester, but not for everyone. This appears to be the insight captured by Williams's snappy dictum: "Life has to have substance if anything is to have sense" (Williams, 1981c, p. 18).[1]

That being said, the transcendental argument, in any of its forms mentioned above, even if sound, would not do much by itself to limit morality's demands. What is established practically is that we cannot make moral demands on people whose lives are at risk of losing all meaning:

> When we focus on someone whose life is so devoid of meaning that she sees no reason to live, it does seem absurd to say to her that she should go out there and maximize utility. . .To the suggestion that she should get up and make the world a better place, t*he deeply depressed person* might well reply, 'Why should I? Why am I responsible for the well-being of the world or the whales?'. . . So far as I can tell, the moralist has nothing to say in answer. (Wolf, 1997, p. 307, my emphasis)

The transcendental argument yields no reason why a person should not be subject to moral demands up to the point of becoming "deeply depressed". We have not vindicated the playing of video games in one's free time unless that is among a person's last remaining meaningful projects. Furthermore, notice that asking from a severely depressed person to maximize utility is just a straightforwardly cruel thing to do, at the very least because it is endangering her mental health and her life. Making demands from the severely depressed is a significant moral cost—but then it is not clear that the transcendental argument challenges Singer's claim that one should act for a moral cause unless this entails sacrificing something of moral significance.

[1] Williams's transcendental argument appears to have its roots in Nietzsche's (2003) idea that contemporary morality is life-denying: it suppresses human life, thereby rooting out the conditions for its own existence (Williams, 1995c, p. 245).

Closely connected to the transcendental argument is the argument from non-moral values suggested by Williams (1981b, pp. 37-38; see also §5.4) and influentially developed by Wolf (1982) in "Moral Saints". The idea of the supererogatory is needed to provide the space for the "nonmoral virtues, as well as many of the interests and personal characteristics that we generally think contribute to a healthy, well-rounded, richly developed character" (Wolf, 1982, p. 421). There are many acceptable personal ideals besides the moral saint, such as "Katharine Hepburn's grace, Paul Newman's 'cool'. . . the high-spirited passionate nature of Natasha Rostov. . . the keen perceptiveness of Lambert Strether" (Wolf, 1982, p. 422). Some of these cannot be pursued if one is supposed to give up all of one's excess resources to morality, as "no plausible argument can justify the use of human resources involved in producing a *pate de canard en croute* against possible alternative beneficent ends to which these resources might be put" (Wolf, 1982, p. 422). Moreover, according to Wolf, some nonmoral virtues are just incompatible with a thoroughgoing moral character. The cynical sense of humor of a Marx Brothers film or a play by George Bernard Shaw cannot coexist with the moral requirement to see everyone in their best light (Wolf, 1982, p. 422). If morality was unconditionally demanding, human life would be duller and emptier because it would be deprived of the diversity of valuable activities and achievements that give it much of its substance.

Although Wolf appears right that morality cannot be maximally overriding, it is again important to see the limited practical consequences of her argument. First, although moral demands cannot be fanatically absolutist, it is still an open question how much of the non-moral interests, achievements, and virtues ought to be sacrificed for morality. Singer's point is precisely to challenge the relative value of yet another modern art exhibition, another *pate de canard en croute,* or of someone's improving their backhand, when the world is plagued by preventable famine, war, disease, and climate change. Second, it is worth noting that moral projects often greatly promote non-moral interests and values too. In fact, this is often part of their goal. If we truly care about beauty, for instance, should we not do our best to protect the planet's endangered species?[2]

Third, and most importantly, Wolf puts into question the *scope* of what we should care about but not the very *requirement* to devote one's personal life to

[2] On questioning the strict divide between moral and aesthetic value, see Sophie Grace Chappell (2018).

something worthy. Thus, taking Wolf's account strictly, I will only be excused from moral demands if I aim to realize some sort of non-moral virtue or achievement instead. The dull-witted and mediocrely talented of us, the ones who cannot hope to have anything close to the "cool" of Paul Newman, would have less justification to escape morality's grip. It would be permissible to play video games only if it would help, say, develop a well-rounded character. However, aiming to create something of value is not why we do many, maybe most, of the things we do in our leisure time. The transcendental argument and the argument from nonmoral values share a flaw: they aim to demonstrate that a person may evade moral demands only by an appeal to something else of (objective) worth, such as a nonmoral virtue or meaningful project. Much like an attempt to escape military subscription, an agent is expected to convince morality that she will do something valuable with her personal time and resources in order to be excused from its demands.

This cannot be right. Missing from this picture is a value that can be called *liberty,* or *autonomy,* or perhaps *integrity* (T. Chappell, 2007, p. 259). A central concern in human life is to secure the freedom to pursue one's own desires.[3] Within certain constraints, such as not harming others, we should be able to do whatever we want, whether it ends up being a good for the world or completely worthless instead. We want to loiter, fool around, stare at the sky, solve crossword puzzles, play mind-numbing video games, build things and then break them apart again, or just spend time doing nothing other than enjoying that no one else's will is burdening us down. My point is that we perform many activities not because they give our lives meaning but simply because we want to. Having autonomy is not necessarily a categorical desire in Williams's sense. If I am to live, I want it to be reasonably autonomously, but I do not necessarily live for the sake of acting autonomously (though I may live for the sake of autonomously *pursuing a particular project)*. If I have leisure time, I want to spend it doing whatever comes to my mind, but I do not necessarily live for the sake of having leisure time. Nor do we do many of the things we do because they help develop a nonmoral virtue or worthy interest. We want to do things

[3] Williams has something close to this value in mind when he says that moral demands that are completely disconnected from the agent's concerns and projects are, "in the most literal sense, an attack on his integrity" (Williams & Smart, 1973, p. 117). In a footnote, he likens Christian humility to the humility required by a secular utilitarianism, where subservience to God is substituted with "subservience to other men and their projects" (Williams & Smart, 1973, p. 117, fn. 1).

even if they are pointless, silly, experimental, self-destructive, trivial, or ephemeral. Without this freedom, we would feel burdened and, at the limits, dominated.

Of course, I am not claiming that there is an absolute and inviolable right to autonomy. My point is merely that autonomy is a value that has *some* weight when set against the moral demands of others, and so there are circumstances when it is permissible to choose the former over the latter. Moreover, there is no *necessary* conflict between autonomy and the moral demands of others: an agent may wish to exercise his autonomy by acting for a moral cause. But sometimes, and I think quite often, a person is faced with the difficult choice between moral demands on the one hand, and, on the other, not meaningful or valuable projects, but simply one's plain wish *to have a personal life*.

6.3 The moral/personal dilemma

Sometimes an appeal to one's autonomy may trump moral demands and vice versa. There is also a grey area where the two considerations have roughly equal weight, and the agent is free to choose either way.[4] On the dominant view of supererogation, my argument should stop here. However, the agent's responsibility following choices in this grey area calls for closer inspection because it takes an exceptional form, especially regarding the moral sentiments.

To start, we must distinguish between two kinds of supererogatory acts, which I shall call *amelioratory* and *additive*. An additive supererogatory act adds something of moral value to the world but there is no noticeable moral cost if it is not performed. Examples include giving gifts to one's coworkers, cleaning up the local park, or walking along with a stranger who asks for directions to show her the way. These are nice things to do, but it would not be particularly bad if they were foregone. In contrast, failure to perform an amelioratory supererogatory act incurs a substantial moral cost to the world. These acts aim at ameliorating a bad state of affairs, such as someone's death or illness, or more globally, famine and climate change.

[4] The moral/personal dilemma is therefore different from being blameworthy and yet justified in a broader sense (§5.4). Gauguin might have a reason to choose to be blameworthy; in contrast, as I will argue in this section, the agent who chooses her personal life over certain supererogatory demands is not blameworthy.

There are no negative moral consequences for an agent if she does not perform additive supererogatory acts. The case is different, however, with amelioratory supererogatory acts, as the agent faces ambivalence: she must make the choice between justifiably acting within the sphere of her autonomy on the one hand and performing an amelioratory supererogatory act on the other. Think of the choice between having a luxurious dinner and donating the money to a worthy cause. Each option may come with a cost. A person choosing to donate may feel sad and frustrated that she is giving up something nice that is within her right to enjoy. However, deciding in favor of the luxurious meal may bring about fitting feelings of guilt and disappointment, accompanying the evaluative judgment that one is falling short of a moral standard. I shall call this kind of double bind the *moral/personal dilemma*.[5]

To understand the sense in which a person fails morally when she chooses her personal life over amelioratory supererogatory demands, and the guilt and disappointment connected to this, it would be instructive to contrast the moral/personal dilemma with two other related phenomena: the classic moral dilemma and innocently possessing a moral debt (explored in §4.5).

Take the classic moral dilemma first. It occurs when all available options breach a moral requirement, and so the agent is blameworthy whatever she does (Greenspan, 1983b; Holbo, 2002; Stocker, 1971; Tessman, 2015; Williams, 1973a). This kind of dilemma, it is argued, leaves the agent with *moral residue,* which is usually spelled out in terms of negative emotions. For Williams, moral residue is constituted by agent-regret (Williams, 1973a; 1981b, p. 31), and other emotions that have been proposed include distress, grief, and, most pertinently to responsibility, guilt (Greenspan, 1983a; Holbo, 2002, p. 266; Tessman, 2015, p. 12). It is worth nothing that these negative emotions are not merited because the agent is *unsure* that what he did was wrong: "[Agamemnon] lies awake, not because of a doubt, but because of a certainty" (Williams, 1973a, p. 173).

[5] Some might find the term "moral/personal dilemma" confusing since it arises out of a conflict between two moral values: autonomy and beneficence (or something like it). However, the agent's *reasons for action* in the alternatives of the dilemma are, at least paradigmatically, respectively non-moral and moral in nature, as for instance when one has to choose between having a luxurious dinner and donating the money to charity.

Let us return to ***Sophie,*** a classic example of the classic moral dilemma, discussed in §4.5. Here is the case again:

> ***Sophie.*** Sophie and her two children are prisoners in a Nazi concentration camp. Sophie's captors make her choose which one of her two children is to be killed; if she refuses to make a decision, they threaten to kill both. Sophie chooses the younger sibling and feels profound guilt even though she knows that there was nothing she or anyone else in her situation could have done better.

According to Patricia Greenspan, Sophie's guilt is fitting; it is "true to the facts" (Greenspan, 1983a, p. 122). The facts are that Sophie is blameworthy for the killing of her child. She is not only "cooperating in an evil human scheme, even if only to limit its effects" (Greenspan, 1983a, p. 122), but her "choice is to *prescribe* death for one of her children" (Greenspan, 1983a, p. 123, emphasis in original). Sophie (like Agamemnon) finds herself in a situation where she cannot avoid becoming a "bad person" (Holbo, 2002, p. 266), as all possible options for action are morally prohibited.

Defenders of moral dilemmas like Greenspan and Holbo rely on what can be called a Miasma View of blameworthiness (Holbo, 2002, p. 266), on which an agent is polluted (made blameworthy) by morally bad acts even if she does not exhibit any character flaws—and so all available actions for Sophie are prohibited, i.e., merit blame. The Miasma View must be assumed to make the case that Sophie faces a moral dilemma, since on other, more standard views of blameworthiness, an agent's action must reveal something flawed about her moral character or her quality of will in order for her to be blameworthy (e.g., Fischer & Ravizza, 1998; McKenna, 2012; Shoemaker, 2015; R. J. Wallace, 1994). Sophie would not be blameworthy on these more standard views of responsibility, as her actions stem from admirable traits, like the strength and resolve to do the best out of unspeakably horrible situations. It is a stretch of ordinary language to say that Sophie is *cooperating* with an evil human scheme by prescribing the death of her child, thereby becoming a bad person. A more accurate description is that Sophie is *coerced* by an evil human scheme. Sophie's rational self-evaluation must include the thought that she could not have done better, nor could have anyone else in her situation. These judgments take away the rational basis for deeming her a bad person—if this is naturally understood as a person with a bad moral character. The only way to make sense of her

blameworthiness is to argue that it attaches not to character flaws, but to performing morally bad acts.

However, the Miasma View, even if it were true, does not help with making sense of the moral/personal dilemma. This is because, as I argued, the agent who chooses her personal life has a *right* to do so; she is not breaching a moral requirement. My description of the moral/personal dilemma relies on the intuition that an appeal to autonomy makes it *permissible* to sometimes choose one's personal life over amelioratory supererogatory demands.

An alternative for making sense of ***Sophie*** that I suggested in section §4.5 is that fitting guilt does not necessarily entail blameworthiness. On a Moral Debt Account of guilt, Sophie is justified all things considered, but her guilt marks a moral debt that she has in virtue of a disrupted relationship. Similarly, the agent who decides in favor of his personal life over moral demands has a moral debt, though he is justified all things considered. He is justified by his right to autonomy, but his guilt is nevertheless an acknowledgment that he owes others his active compassion.

However, even on this reading, the moral/personal dilemma cannot be read as strictly analogous to ***Sophie***. In contrast to Sophie, the agent who justifiably decides in favor of his personal life over moral demands *fails along a moral dimension* because he can perform a morally better act but decides not to. Such an agent is not coerced; he freely chooses against performing an amelioratory supererogatory act. His action does not merely merit guilt, but also an interpersonal judgment by others that his character is not fully committed to morality; he is failing along a moral dimension.

By what standard does an agent fail morally in the moral/personal dilemma? One possibility is that an agent may feel not that she has failed the moral demands made on her by others, but that she has failed *herself* by not living up to her own self-set moral standards. Failures by one's own norms are often accompanied by a guilty recognition of *akrasia*, as for instance when one cheats on one's diet or exercise routine. (If you are fortunate enough to be unfamiliar with this phenomenon, do an online search of "exercise guilt"). It is possible, then, that an agent feels guilty for not living up to his own personal ideal of someone who answers morality's supererogatory demands.

Yet, while this kind of guilt may be elicited in some, it would not be rational for people who do not aspire to be particularly righteous. An agent may wish to

meet her own moral ideal by performing a supererogatory act but fail due to akrasia; however, she may also reflectively endorse her choice of personal life over morality. She would then have no grounds for guilt about failing her goals because her goals would not include going beyond the call of duty. On this account, then, a person can shake off her guilt simply by deciding to set more modest moral goals for herself. But this fails to capture the objective (or intersubjective) dimension of the moral failure at issue. It seems that it is not up to the agent to choose whether she has fallen short of a moral standard by foregoing amelioratory supererogatory acts.

Telech and Katz (2022) provide a useful framework for approaching the issue in an alternative way: an agent may fail morally, and thereby merit guilt and the disappointment, by shattering others' *normative hopes. Reactive disappointment*, according to Telech and Katz (2022), is a natural blaming response that is evoked when someone fails to live up to another's normative hope, a hope *in* somebody as opposed to a hope *that* something happens. Normative hopes in turn arise out of personal relationships or shared values.[6]

There can be legitimate normative hopes in the agent, based on the shared values of solidarity and compassion, that she chooses a moral cause like helping people in need over her personal life. When the agent does not meet these hopes, she merits reactive disappointment, which typically evokes guilt (Telech & Katz, 2022, p. 873). The agent might imagine what she would have to say to the people who could have benefitted from her moral choice were she to confront them: "I know that I could have helped provide food for your malnourished children today, but I chose not to. Instead, I decided to spend the money on a luxurious meal for myself". It would be strange to be able to say this with a tone of unambivalent affirmation. Even if the agent believes herself justified, guilt is necessary to express the recognition that her choice's grave consequences for

[6] Telech and Katz (2022) say that normative hope *paradigmatically* presupposes "thick personal relationships and shared values, though not necessarily face to face experiences" (p. 865), but later on make the stronger claim that normative hope *necessarily* presupposes such thick relations, and hence cannot be directed at complete strangers (Telech & Katz, 2022, p. 867, fn. 41). I do not see a reason why we should accept the stronger claim. There are certain values, like compassion and beneficence, which we can assume to share with every member of the human moral community. It appears intelligible to be disappointed in humanity in general or in a complete stranger in particular. My hopes and disappointments regarding a stranger who decides not to pull over to help me with my car troubles would be different from those directed at the bad weather.

others disappoint legitimate hopes in her. She is revealed as not compassionate enough.

Guilt in such cases appears to not only be fitting, but a noninstrumentally good response, because it is an acknowledgment of one's moral debt and moral fault (Carlsson, 2017; Clarke, 2013, pp. 155-157; see also §4.5). Conversely, failure to feel guilty (to some appropriate degree) looks like a noninstrumental disvalue. Imagine again someone who can state without guilt to the people in destitute conditions that she chooses to spend her excess time and money on some piece of luxury rather than help them. Caring about others (and nature, animals, etc.) entails feelings of sadness when they are harmed and guilt when this harm flows from one's agency.

6.4 Justification and blame

If guilt and reactive disappointment in the agent is fitting, in what sense can one say that she is justified in her choice by her appeal to autonomy? For Telech and Katz, reactive disappointment is necessarily a blaming response, alongside guilt, resentment and indignation. One may object, therefore, that despite my insistence that autonomy may override some moral demands, the presence of fitting guilt when this happens proves that this is false and the agent in such cases just is blameworthy.

To settle this, we need to be clear about the notion of blameworthiness that is at stake. Let us return again to Gary Watson's (2004b; see also §4.6) distinction between accountability and attributability. Recall that if an agent is accountability-blameworthy, she has performed a moral wrong and in virtue of that is *pro tanto* open to some kind of sanction—which may range from legal punishment to giving someone the cold shoulder—as well as to moral demands, e.g. for an apology, acknowledgment, or reparation. She is also open to the focused negative reactive attitudes of resentment, indignation, and guilt (Watson, 2004b, p. 276). The dominant view is that these accountability responses normatively come together in a package. Guilt, resentment, and indignation are respectively the first-, second-, and third-personal variants of moral anger, and hence if guilt is fitting, so are resentment and indignation (Shoemaker, 2015); and these negative sentiments accompany at least some demands or sanctions that are *pro tanto* justified (e.g. Bennett, 2002; Macnamara, 2011; McKenna, 2021; P. F. Strawson, 2003; R. J. Wallace, 1994; Watson, 2004b).

In contrast, an agent is *attributability-responsible* if she merits an aretaic evaluation, e.g., being lazy or a coward. However, it is "nobody's business" to impose sanctions or to make demands purely in virtue of these agential appraisals (Watson, 2004b, p. 267). Aretaically appraising an individual is necessary, but not sufficient for *holding* her responsible. An essential difference between the two faces of responsibility is that accountability norms, if breached, make the agent liable to some form of focused reactive attitude, sanction or demand, while attributability ones do not.

My account of the moral/personal dilemma serves as a counterexample to the necessary bundling of accountability responses. Neither fitting guilt nor fitting reactive disappointment necessarily entail the *pro tanto* appropriateness of sanctions or demands in the cases I am describing. Although there is an emotional and evaluative cost to the agent's choice, she is nevertheless justifiably acting within the sphere of her autonomy, where this means that she is not liable to unwelcome treatment. Still, others can make the aretaic judgment that he is not a very morally devoted person, accompanied by disappointment.

This is consistent with Telech and Katz's account of reactive disappointment, even though they want to classify the emotion as a form of accountability-blame. On their view, reactive disappointment does not necessarily entail the right to enforce a moral standard or to *demand* an apology, restitution, or acknowledgment. It merely *urges* its addressee "in a backward-looking manner, to e.g. regain what they have lost in their moral failure" (Telech & Katz, 2022, p. 870). We can therefore say such things as "I am disappointed in you, but it is none of my business to demand that you act differently". In fact, we are often reactively disappointed in our close ones without believing at the same time that we can meddle in their lives. Reactive disappointment can thus be disconnected from the rest of the accountability package that includes demands and sanctions (Pereboom, 2014, p. 134).

The appropriateness of individual kinds of accountability responses can thus come apart, and it is in this sense that the agent choosing her personal life over amelioratory supererogatory demands is justified despite her acts' meriting guilt and reactive disappointment. Thus, the personal choice in the moral/personal dilemma is not just an instance of the suberogatory (or of its subset, "failures of common decency" (Calhoun, 2004)), i.e. acts that are bad but not forbidden for which the agent is blameworthy (Driver, 1992). Driver (1992) argues that suberogatory acts are possible when there is a "conflict between an ideal and a

right" (p.287); the moral/personal dilemma seems to fit the bill because it is a case where a right to autonomy conflicts with the moral ideals of solidarity and compassion. Yet, according to Driver, agents are *blameworthy* for suberogatory acts; the suberogatory is needed as a category in the first place is to expand *blameworthiness* to certain kinds of non-forbidden acts (Driver, 1992, p. 287).[7] But unlike Driver and Calhoun, I am not attempting to expand the scope of blameworthiness. Instead, through the case of the moral/personal dilemma, I am questioning the idea that all acts can be separated into either meriting accountability-blame or not.

6.5 Conclusion

I have questioned the neat divide between the supererogatory and the morally required, and the blameworthy and the blameless. Sometimes—and for many in contemporary life, it seems, every day—one must decide between two permissible alternatives. A person can choose her personal life and experience guilt and reactive disappointment, or instead choose morality by sacrificing her own projects and feel sadness and frustration.

One might object that my account overmoralizes and ends up with the same problem that the sphere of the supererogatory was supposed to solve: morality is rendered overly burdensome, even if not strictly speaking overly demanding. After all, on my account, though the agent is justified in sometimes choosing her personal life over moral demands, failure to feel guilty following such a choice is a non-instrumentally bad outcome. This looks remarkably close to a churchy attitude to life wherein one feels guilty for little more than existing.

I admit that this is not an optimal result—and yet this does not by itself mean that it is not true. The objection would only be valid under the hopeful assumption that we can construct a moral theory that imposes a comfortable degree of moral burdens, part of a more general tendency to assume that moral philosophizing should lead to a decrease, rather than an increase of ethical tension. My purpose in this chapter has been to challenge these assumptions. Ordinary experience suggests that we discover moral values and must live with

[7] According to some, the suberogatory is an unnecessary concept in the first place, as it just identifies "impermissible actions that are not rights violations" (Liberto, 2012, p. 398). A more capacious notion of moral wrongness would make the suberogatory redundant.

the emotions and requirements they give rise to, even when they are tragically difficult and messy (S. G. Chappell, 2015).

My account is revisionary in terms of theory but not necessarily in terms of practice. It is possible that we in fact do experience a lot more (fitting) guilt than theories of responsibility acknowledge. Though unfortunate, it is not surprising that we can be overburdened by moral requirements in a hyperconnected modern world. Moral sentiments like guilt plausibly evolved to serve us in contexts far more local and restricted than the global village of today, where each day our actions and inactions affect others all over the planet, sometimes even making the difference between life and death for people who we will never know. This is an unnatural normative landscape for beings who are capable of both compassion and guilt.

Not only is it possible to affect many faraway others through one's daily choices, but it also appears to be the standard predicament for many people around the world. This makes the dilemma between one's personal life and supererogatory actions not a relatively exceptional occurrence, like the difficult choice of Sophie, but rather a cultural and political issue. It introduces yet another inherent problem for a certain kind of liberal outlook, for at its very core lie the ideals of personal autonomy and the equal intrinsic worth of persons. Many liberals believe that the acceptable moral constraints on an individual's behavior should be fairly minimal, and mostly of a negative kind: there are things the agent must not do, but few things that she must do. I have argued that this autonomy comes at a price. When the effects of one's agency extend all over the globe, the ideal of autonomy necessarily conflicts with our natural sympathy and care for the victims of suffering and destruction. The guilt that arises from this conflict is a societal force, and various attempts can be observed in our culture to deal with it by suppressing it, explaining it away, or pretending it is not there. Neither of these is the truthful way to approach the prevalence of guilt; we should rather honestly acknowledge its place in modern life.

Conclusion: responsibility and ambivalence

I mentioned in the Introduction that ambivalence is perhaps an underwhelming response to ethical problems (§1.1.2). It does not solve anything but rather concludes that it is bad whatever one does. I have tried to make the monograph more exciting by spelling out in detail exactly why it is bad. There are many forms of bad and it is of philosophical interest to know the kinds of bad that one can choose between.

However, my goal was not only to characterize a few instances of ambivalence in responsibility. The five debates that I chose can be approached with the concept of ambivalence in order to help with reconciling opposing sides. One way in which philosophical thought progresses is by formulating a new position on an issue that sparks a debate; another, less often encountered but nevertheless necessary, is to aim to resolve a debate by seeking reconciliation between conflicting positions.

Here is, then, a summary of the ambivalences I have identified and the reconciliations I have paved the way for.

In Chapter 2, I argued that there is ambivalence about psychopathic wrongdoing: responding to a psychopath's morally wrong act in unwelcome ways can be, on the one hand, retributively and instrumentally justified, but, on the other, unfair. The reconciliation sought is between those who hold that psychopaths can be blameworthy and those who deny it.

In Chapter 3, I argued that the same kind of ambivalence as the one outlined in Chapter 2 holds for all agents in virtue of their being subject to ultimate luck. The reconciliation sought is between free will sceptics and compatibilists.

In Chapter 4, I argued that there is ambivalence about guilt deserved by morally innocent agents who have a moral debt: it is justified, and yet unfair, to make

them acknowledge their moral debt through guilt. The reconciliation sought is between those who hold that guilt is deserved only by blameworthy agents and those who deny it.

In Chapter 5, I explored the possibility of ambivalence about the choice between becoming blameworthy and sacrificing something else of value, such as one's happiness. If this ambivalence is possible, then so is indirect moral luck: it may be outside of one's control whether one will transform into someone who will regret one's choice of becoming blameworthy or would affirm it instead. The reconciliation sought is between those who deny that outcomes can affect moral evaluations and their opponents.

In Chapter 6, I argued that an agent may feel ambivalent about the choice between her personal life and certain supererogatory demands. Such an agent, though she is not blameworthy, would be burdened by fitting guilt and disappointment if she chooses the former over the latter. The reconciliation sought is between those who defend a more demanding and those who defend a less demanding theory of morality.

The broader picture of responsibility that emerges from my accounts of ambivalence can be better appreciated if contrasted with a philosophical outlook exemplified by some forms of ancient Stoicism. Some Stoic thinkers, impressed by the order of the world, concluded that the universe must be unfolding in accordance with a rational plan. Every event, even the ones that from the human perspective are experienced as irritating, bad, or evil, serves some function and perfectly fits with everything else, like a piece in a jigsaw puzzle (Annas, 1993, p. 161; Long, 1971).

Accordingly, the Stoic sage sees the grand scheme of things and exemplifies a masculine warrior ideal: he is someone who can lead a happy life despite all the seemingly horrible things that can happen to him in his life. Theophrastus, who claimed that "Fortune, not wisdom, rules life", was mocked by Stoics because of his "weak-kneed, contemptible response to pain and misfortune" (Annas, 1993, p. 388).

While many contemporary philosophers do not subscribe to the Stoic worldview, some nevertheless share with them an assumption that the elements of our ethical lives are, or can be made to be, in rational harmony with each other. I have shown that, quite on the contrary, responsibility is replete with incongruities and tragic trade-offs. Responsibility responses do not come

together in a well-ordered package: guilt can be fitting when resentment is not; making demands of psychopaths can be pointless despite their blameworthiness; much of our blaming practices are ultimately unfair, and one's perspective on one's own moral responsibility changes along with one's evaluative transformations. While evolution has indeed supplied us with emotions and intuitions that are remarkably well-suited to the challenges we face as human beings, it would be too quick to conclude that these emotions and intuitions seamlessly fit with each other.

In the beginning of the monograph, I suggested that some tensions in moral responsibility comprise a subset of existential problems. Now, after I have presented these tensions in detail, I hope that the nature of these problems of ambivalence has become clearer. If, as the Stoics supposed, everything fit together in accordance with a rational plan, then existential problems would be overcome by repressing one's illusory concerns, for instance, by rejecting one's fear of death as a mere weakness in one's psychology. Existential problems would then be resolvable by pure strength of character. Problems of ambivalence in responsibility, however, are such that they entail a necessary ethical cost. When faced with these problems, no strength of character can make it so that there is no sacrifice of ethical value.

References

Adams, R. M. (1985). Involuntary Sins. *The Philosophical Review*, *94*(1), 3-31. https://doi.org/10.2307/2184713

Aharoni, E., Antonenko, O., & Kiehl, K. A. (2011). Disparities in the moral intuitions of criminal offenders: The role of psychopathy. *Journal of Research in Personality*, *45*(3), 322-327. https://doi.org/10.1016/j.jrp.2011.02.005

Aharoni, E., & Fridlund, A. I. (2013). Moralistic Punishment as a Crude Social Insurance Plan. In T. A. Nadelhoffer (Ed.), *The Future of Punishment* (pp. 213-229). Oxford University Press.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Association.

Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). "False positive" emotions, responsibility, and moral character. *Cognition*, *214*, 104770. https://doi.org/10.1016/j.cognition.2021.104770

Andersson, H., & Herlitz, A. (2021). Introduction. In H. Andersson & A. Herlitz (Eds.), *Value Incommensurability: Ethics, Risk, and Decision-Making* (pp. 1-27). Routledge.

Andersson, H., & Herlitz, A. (2022). Classifying comparability problems in a way that matters. *Synthese*, *200*(4), 322. https://doi.org/10.1007/s11229-022-03795-8

Annas, J. (1993). *The Morality of Happiness*. Oxford University Press.

Baras, D., & Na'Aman, O. (2022). What Makes Something Surprising? *Philosophy and Phenomenological Research*, *105*(1), 195-215.

Baumard, N. (2016). *The Origins of Fairness: How Evolution Explains Our Moral Nature*. Oxford University Press.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243-267. https://doi.org/10.1037/0033-2909.115.2.243

Baumeister, R. F., Wotman, S. R., & Stillwell, A. M. (1993). Unrequited love: On heartbreak, anger, guilt, scriptlessness, and humiliation. *Journal of Personality and Social Psychology*, *64*, 377-394. https://doi.org/10.1037/0022-3514.64.3.377

Bennett, C. (2002). The varieties of retributive experience. *Philosophical Quarterly*, *52*(207), 145-163.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, *57*(1), 1-29.

Blair, R. J. R., & Cipolotti, L. (2000). Impaired social response reversal. A case of 'acquired sociopathy'. *Brain*, *123 (Pt 6)*, 1122-1141. https://doi.org/10.1093/brain/123.6.1122

Blair, R. J. R., Hwang, S., White, S. F., & Meffert, H. (2016). Emotional Learning, Psychopathy, and Norm Development. In S. M. Liao (Ed.), *Moral Brains: The Neuroscience of Morality* (pp. 185-202). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199357666.003.0009

Bok, H. (1998). *Freedom and Responsibility*. Princeton University Press.

Brink, D. O., & Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility* (Vol. 1, pp. 284-313). Oxford University Press.

Broome, J. (1997). Is Incommensurability Vagueness? In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press.

Brosnan, S., & de Waal, F. (2003). Monkeys Reject Unequal Pay. *Nature*, *425*(6955), 297. https://doi.org/10.1038/nature01963

Calhoun, C. (2004). Common decency. In C. Calhoun (Ed.), *Setting the Moral Compass: Essays by Women Philosophers* (pp. 128--142). Oxford University Press.

Carlsson, A. B. (2017). Blameworthiness as Deserved Guilt. *The Journal of Ethics*, *21*(1), 89-115.

Chang, R. (2002). The Possibility of Parity. *Ethics*, *112*(4), 659-688. https://doi.org/10.1086/339673

Chappell, S. G. (2015). How encounters with values generate demandingness. In M. van Ackeren & M. Kuehler (Eds.), *The Limits of Obligation* (pp. 84-99). Routledge.

Chappell, S. G. (2018). Duty, beauty and booty: an essay in ethical reappropriation. In M. van Ackeren & S. G. Chappell (Eds.), *Ethics Beyond the Limits: New Essays on Bernard Williams' Ethics and the Limits of Philosophy* (pp. 94-118). Routledge.

Chappell, S. G. (2022). *Epiphanies: An Ethics of Experience*. Oxford University Press.

Chappell, T. (2007). Integrity and Demandingness. *Ethical Theory and Moral Practice*, *10*(3), 255-265.

Chisholm, R. M. (1976). *Person and Object: A Metaphysical Study*. Open Court.

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

Clarke, R. (2013). Some Theses on Desert. *Philosophical Explorations*, *16*(2), 153-164.
Clarke, R. (2016). Moral Responsibility, Guilt, and Retributivism. *The Journal of Ethics*, *20*(1/3), 121-137.
Coates, D. J. (2023). *In Praise of Ambivalence*. Oxford University Press.
Cohen, G. A. (2000). If You're an Egalitarian, How Come You're so Rich? *The Journal of Ethics*, *4*(1/2), 1-26.
Cohen, J., & Greene, J. (2006). For the Law, Neuroscience Changes Nothing and Everything. In S. Zeki & O. Goodenough (Eds.), *Law and the Brain*. Oxford University Press.
Cuypers, S. E. (2013). Moral Shallowness, Metaphysical Megalomania, and Compatibilist-Fatalism. *Ethical Theory and Moral Practice*, *16*(1), 173-188.
D'Arms, J., & Jacobson, D. (1994). Expressivism, Morality, and the Emotions. *Ethics*, *104*(4), 739-763.
D'Arms, J., & Jacobson, D. (2000). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophical and Phenomenological Research*, *61*(1), 65-90.
D'Arms, J., & Jacobson, D. (2022). The Motivational Theory of Guilt (and Its Implications for Responsibility). In A. B. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 11-27). Cambridge University Press. https://doi.org/DOI: 10.1017/9781009179263.002
D'Arms, J., & Jacobson, D. (2023). *Rational Sentimentalism*. Oxford University Press. https://doi.org/10.1093/oso/9780199256402.003.0002
Darwall, S. (2006). *The Second Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.
Debove, S., Baumard, N., & André, J. (2017). On the evolutionary origins of equity. *PLOS ONE*, *12*(3).
Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.
Deonna, J., & Teroni, F. (2008). Differentiating Shame from Guilt. *Consciousness and Cognition*, *17*(4), 1063-1400.
Deonna, J., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction*. Routledge.
Deonna, J., & Teroni, F. (2015). Emotions as Attitudes. *Dialectica*, *69*(3), 293-311.
Deonna, J., & Teroni, F. (2022). Emotions and Their Correctness Conditions: A Defense of Attitudinalism. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00522-0
Dill, B., & Darwall, S. (2014). Moral psychology as accountability. In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 40-83). Oxford University Press.
Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy*, *70*(3), 286 – 295.

Duggan, A. P. (2018). Moral Responsibility as Guiltworthiness. *Ethical Theory and Moral Practice*, *21*(2), 291-309. https://doi.org/10.1007/s10677-018-9863-0

Dworkin, R. (2000). *Sovereign Virtue*. Harvard University Press.

Echeverri, S. (2019). Emotional Justification. *Philosophy and Phenomenological Research*, *98*(3), 541-566.

Ekstrom, L. W. (1999). *Free Will: A Philosophical Study*. Westview Press.

Enoch, D. (2012). Being Responsible, Taking Responsibility, and Penumbral Agency. In U. Heuer & G. Lang (Eds.), *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams* (pp. 95-132). Oxford University Press.

Enoch, D., & Marmor, A. (2007). The case against moral luck. *Law and philosophy*, *26*(4), 405-436.

Feinberg, J. (1973). *Social Philosophy*. Prentice-Hall.

Fischer, J. M. (2006). The Cards that are Dealt You. *Journal of Ethics*, *10*(1-2), 107-129.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Foot, P. (1978). Hume on Moral Judgment. In *Virtues and Vices* (pp. 74-80). Oxford University Press.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*(1), 5-20.

Fricker, M. (2022). Bernard Williams as a Philosopher of Ethical Freedom. In M. Talbert & A. Szigeti (Eds.), *Morality and Agency: Themes from Bernard Williams* (pp. 265-289). Oxford University Press.

Frijda, N. H. (1986). *The emotions*. Editions de la Maison des Sciences de l'Homme.

Glenn, A. L., Iyer, R., Graham, J., Koleva, S., & Haidt, J. (2009). Are all types of morality compromised in psychopathy? *Journal of Personality Disorders*, *23*(4), 384-398. https://doi.org/10.1521/pedi.2009.23.4.384

Greenspan, P. S. (1983a). Moral dilemmas and guilt. *Philosophical Studies*, *43*(1), 117 - 125.

Greenspan, P. S. (1983b). Moral dilemmas and guilt. *Philosophical Studies*, *43*(1), 117-125.

Griffith, M. (2010). Why agent-caused actions are not lucky. *American Philosophical Quarterly*, *47*(1), 43-56.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press.

Hieronymi, P. (2004). The Force and Fairness of Blame. *Philosophical Perspectives*, *18*, 115-148.

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, *161*(3), 357-373.

Holbo, J. (2002). Moral Dilemmas and the Logic of Obligation. *American Philosophical Quarterly*, *39*(3), 259 - 274.

Holmes, O. W. (2012). *The Essential Holmes: Selections from the Letters, Speeches, Judicial Opinions, and Other Writings of Oliver Wendell Holmes, Jr* (R. A. Posner, Ed.). University of Chicago Press. https://books.google.se/books?id=m3WQDwAAQBAJ

Honderich, T. (1988). *A Theory of Determinism: The Mind, Neuroscience, and Life-hopes*. Oxford University Press.

Hume, D. (2009). *A Treatise of Human Nature*. The Floating Press. (1740)

Hutcheson, F. (2008). *Inquiry into the Original of Our Ideas of Beauty and Virtue.* (W. Leidhold, Ed.). Liberty Fund.

Inbar, Y., Pizarro, D. A., Gilovich, T., & Ariely, D. (2013). Moral masochism: on the connection between guilt and self-punishment. *Emotion*, *13*(1), 14-18. https://doi.org/10.1037/a0029749

Jacobson, D. (2013). Regret, Agency, and Error. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility (Volume 1)*. Oxford University Press.

Jacobson, D., & D'Arms, J. (2006). Anthropocentric Constraints on Human Value. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 1, pp. 99-126). Oxford University Press.

Kamtekar, R., & Nichols, S. (2019). Agent-Regret and Accidental Agency. *Midwest Studies in Philosophy*, *43*(1), 181-202.

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

Kant, I. (1998). *Groundwork of the Metaphysics of Morals* (M. J. Gregor, Ed. & Trans.). Cambridge University Press. (1785)

Kant, I. (2017). *The Metaphysics of Morals* (M. Gregor, Trans.; L. Denis, Ed. 2 ed.). Cambridge University Press. (1797)

Kavka, J. (1949). Pinel's Conception of the Psychopathic State: An Historical Critique. *Bulletin of the History of Medicine*, *23*(5), 461-468.

Kerr, A. D. (2019). Anticipatory Guilt. In B. Cokelet & C. J. Maley (Eds.), *The Moral Psychology of Guilt* (pp. 182-210). Rowman & Littlefield International.

Lamb, R. E. (1983). Guilt, shame, and morality. *Philosophy and Phenomenological Research*, *43*(3), 329-346.

Lang, G. (2018). Gauguin's Lucky Escape: Moral Luck and the Morality System. In S. G. Chappell & M. v. Ackeren (Eds.), *Ethics Beyond the Limits: New Essays on Bernard Williams' Ethics and the Limits of Philosophy* Routledge.

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

Larmore, C. (1999). The Idea of a Life Plan. *Social Philosophy and Policy*, *16*(1), 96-112.

Latus, A. (2001). Moral luck. In J. Feiser (Ed.), *Internet Encyclopedia of Philosophy*.
Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, and Psychology*, *14*(2), 129-138.
Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.
Levy, N. (2014). Psychopaths and blame: The argument from content. *Philosophical Psychology*, *27*(3), 351–367.
Liberto, H. R. (2012). Denying the Suberogatory. *Philosophia*, *40*(2), 395-402.
Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, *106*, 623-642. https://doi.org/10.1093/brain/106.3.623
Long, A. A. (1971). The Logical Basis of Stoic Ethics. *Proceedings of the Aristotelian Society*, *71*(1), 85-104.
Lowe, E. J. (2008). *Personal agency: the metaphysics of mind and action*. Oxford University Press.
Lukomska, A. (forthcoming). The Moral Significance of Agent-Regret. *Journal of Philosophical Research*.
Mackie, J. L. (1982). Morality and the Retributive Emotions. *Criminal Justice Ethics*, *1*(1), 3-10.
Macnamara, C. (2011). Holding others responsible. *Philosophical Studies*, *152*(1), 81-102.
Mason, E. (2019). *Ways to Be Blameworthy: Rightness, Wrongness, and Responsibility*. Oxford University Press.
McCann, H. J. (1998). *The Works of Agency: On Human Action, Will and Freedom*. Cornell University Press.
McCloskey, H. J. (1957). An examination of restricted utilitarianism. *Philosophical Review*, *66*(4), 466-485.
McKenna, M. (2012). *Conversation & Responsibility*. Oxford University Press.
McKenna, M. (2020). Punishment and the value of deserved suffering. *Public Affairs Quarterly*, *34*(2), 97-123. https://doi.org/10.2307/26921122
McKenna, M. (2021). Wimpy Retributivism and the Promise of Moral Influence Theorists. *The Monist*, *104*(4), 510-525.
McKenna, M. (2022). Guilt and Self-Blame within a Conversational Theory of Moral Responsibility. In A. B. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 151-174). Cambridge University Press. https://doi.org/DOI: 10.1017/9781009179263.009
McKenna, M., & Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. Routledge.

Metz, T. (2018). Survivor's Guilt. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (pp. 1-8). Wiley.

Milona, M. (2016). Taking the Perceptual Analogy Seriously. *Ethical Theory and Moral Practice*, *19*(4), 897-915. https://doi.org/10.1007/s10677-016-9716-7

Mitchell, J. (2021). *Emotion as Feeling Towards Value: A Theory of Emotional Experience*. Oxford University Press. https://doi.org/10.1093/oso/9780192846013.001.0001

Moller, D. (2007). Love and death. *Journal of Philosophy*, *104*(6), 301-316.

Morse, S. I. (2013). Compatibilist Criminal Law. In T. A. Nadelhoffer (Ed.), *The Future of Punishment* (pp. 107-131). Oxford University Press.

Morse, S. J. (2013). Psychopathy and the law: the United States experience. In L. Malatesti & J. McMillan (Eds.), *Responsibility and Psychopathy: Interfacing law, psychiatry and philosophy* (pp. 41-62). Oxford University Press.

Mounce, H. O. (1999). *Hume's Naturalism*. Routledge.

Mulligan, K. (2007). Intentionality, Knowledge and Formal Objects. *Disputatio*, *2*(23), 205-228. https://doi.org/doi:10.2478/disp-2007-0010

Murphy, L. B. (2000). *Moral Demands in Nonideal Theory*. Oxford University Press.

Müller, J. M. (2017). How (Not) to Think of Emotions as Evaluative Attitudes. *Dialectica*, *71*(2), 281-308.

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

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Nelissen, R. M. A. (2012). Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science*, *3*(2), 139-144. https://doi.org/10.1177/1948550611411520

Nelkin, D. K. (2015). Psychopaths, Incorrigible Racists, and the Faces of Responsibility. *Ethics*, *125*(2), 357-390.

Nelkin, D. K. (2021). Moral Luck. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2021 Summer Edition ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/moral-luck/

Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. Oxford University Press.

Nietzsche, F. (2003). *The Genealogy of Morals* (H. B. Samuel, Trans.; T. N. R. Rogers, Ed.). Dover Publications.

Nussbaum, M. C. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

Nussbaum, M. C. (2004). *Hiding From Humanity: Disgust, Shame, and the Law*. Princeton University Press.

O'Shaughnessy, B. (1980). *The Will: A Dual Aspect Theory*. Cambridge University Press.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.

Palmer, D. (2020). Free will and control: a noncausal approach. *Synthese*, *198*(10), 10043-10062.

Pardo, M. S., & Patterson, D. (2013). Neuroscience, Normativity, and Retributivism. In T. A. Nadelhoffer (Ed.), *The Future of Punishment* (pp. 133-153). Oxford University Press.

Paul, L. A. (2014). *Transformative Experience*. Oxford University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pereboom, D. (2017). Honderich on Freedom, Determinism, and Meaning in Life. In G. Caruso (Ed.), *Ted Honderich on Consciousness, Determinism, and Humanity* (pp. 143-158). Palgrave Macmillan.

Pereboom, D. (2021). *Wrongdoing and the Moral Emotions*. Oxford University Press.

Portmore, D. W. (2022). A Comprehensive Account of Blame: Self-Blame, Non-Moral Blame, and Blame for the Non-Voluntary. In A. B. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 48-76). Cambridge University Press.

Prinz, J. J., & Nichols, S. (2010). Moral emotions. In J. M. Doris (Ed.), *Moral Psychology Handbook* (pp. 111-146). Oxford University Press.

Proctor, D., Williamson, R. A., de Waal, F. B. M., & Brosnan, S. F. (2013). Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences*, *110*(6), 2070-2075. https://doi.org/10.1073/pnas.1220806110

Queloz, M. (2021a). The Essential Superficiality of the Voluntary and the Moralization of Psychology. *Philosophical Studies*, *179*, 1591–1620.

Queloz, M. (2021b). *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*. Oxford University Press.

Queloz, M. (2022). A Shelter from Luck: The Morality System Reconstructed. In A. Szigeti & M. Talbert (Eds.), *Morality and Agency: Themes from Bernard Williams* (pp. 182-209). Oxford University Press.

Rabinowicz, W. (2009). Incommensurability and Vagueness. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *83*, 71-94.

Rabinowicz, W., & Rønnow-Rasmussen, T. (2004). The strike of the demon: On fitting pro-attitudes and value. *Ethics*, *114*(3), 391-423.

Range, F., Horn, L., Viranyi, Z., & Huber, L. (2009). The absence of reward induces inequity aversion in dogs. *Proceedings of the National Academy of Sciences*, *106*(1), 340-345. https://doi.org/10.1073/pnas.0810957105

Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

Richards, D. A. J. (1971). *A Theory of Reasons for Action*. Clarendon Press.

Roberts, R. C. (2003). *Emotions: An Essay in Aid of Moral Psychology*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511610202

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, *18*(1), 295–313.

Russell, P. (2017). Free Will Pessimism. In *The Limits of Free Will: Selected Essays* (pp. 243-275). Oxford University Press.

Russell, P. (2022). Free Will and the Tragic Predicament: Making Sense of Williams. In M. Talbert & A. Szigeti (Eds.), *Morality and Agency: Themes from Bernard Williams* (pp. 161-181). Ofxord University Press.

Sanz-García, A., Gesteira, C., Sanz, J., & García-Vera, M. P. (2021). Prevalence of Psychopathy in the General Adult Population: A Systematic Review and Meta-Analysis [Systematic Review]. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.661044

Scanlon, T. (1998). *What We Owe to Each Other*. Harvard University Press.

Sellars, W. S. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Science, Perception, and Reality* (pp. 35-78). Humanities Press/Ridgeview.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, *121*(3), 602-632.

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press. https://books.google.se/books?id=hDN2BwAAQBAJ

Shoemaker, D. (2019). Hurt Feelings. *Journal of Philosophy*, *116*(3), 125-148.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, *1*(3), 229-243.

Smart, J. J. C. (1961). Free will, praise and blame. *Mind*, *70*(279), 291-306.

Smilansky, S. (2000). *Free Will and Illusion*. Oxford University Press.

Smilansky, S. (2011). Hard Determinism and Punishment: A Practical Reductio. *Law and philosophy*, *30*(3), 353-367.

Smilansky, S. (2017). Pereboom on Punishment: Funishment, Innocence, Motivation, and Other Difficulties. *Criminal Law and Philosophy*, *11*(3), 591-603.

Smith, A. M. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*, *115*(2), 236-271.

Smith, A. M. (2015). Responsibility as Answerability. *Inquiry: An Interdisciplinary Journal of Philosophy*, *58*(2), 99-126.

Sommers, T. (2012). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton University Press.

Stocker, M. (1971). 'Ought' and 'can'. *Australasian Journal of Philosophy*, *49*(3), 303 – 316.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies*, *75*(1-2), 5-24.

Strawson, P. F. (1985). *Skepticism and Naturalism: Some Varieties*. Routledge.

Strawson, P. F. (2003). Freedom and Resentment. In G. Watson (Ed.), *Free Will* (2 ed., pp. 72-93). Oxford University Press.

Styron, W. (1979). *Sophie's Choice*. Random House.

Sussman, D. (2018). Is Agent-Regret Rational? *Ethics*, *128*(4), 788-808.

Szigeti, A. (2015). Sentimentalism and Moral Dilemmas. *Dialectica*, *69*(1), 1-22.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy? *Pacific Philosophical Quarterly*, *89*(4), 516-535.

Talbert, M. (2012). Moral Competence, Moral Blame, and Protest. *The Journal of Ethics*, *16*(1), 89-109.

Talbert, M. (2014). The Significance of Psychopathic Wrongdoing. In T. Schramme (Ed.), *Being Amoral: Psychopathy and Moral Incapacity*. MIT Press.

Talbert, M. (2021). Psychopaths and Symmetry: A Reply to Nelkin. *Philosophia*, *49*, 1233–1245.

Tappolet, C. (2016). *Emotions, Values, and Agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199696512.001.0001

Taylor, C. (1976). Responsibility for self. In A. O. Rorty (Ed.), *The Identities of Persons* (pp. 281--299). University of California Press.

Taylor, G. (1985). *Pride, Shame, and Guilt: Emotions of Self-Assessment*. Oxford University Press.

Telech, D. (2022). Relation-Regret and Associative Luck. In A. Szigeti & M. Talbert (Eds.), *Agency, Fate and Luck: Themes from Bernard Williams* (pp. 236-267). Oxford University Press.

Telech, D., & Katz, L. D. (2022). Condemnatory Disappointment. *Ethics*, *132*(4), 851-880.

Temkin, L. (2017). Equality as Comparative Fairness [Article]. *Journal of Applied Philosophy*, *34*(1), 43-60. https://doi.org/10.1111/japp.12140

Tessman, L. (2015). *Moral Failure: On the Impossible Demands of Morality*. Oxford University Press.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford University Press.

Vargas, M., & Nichols, S. (2007). How to be fair to psychopaths. *Philosophy, Psychiatry, and Psychology*, *14*(2), 153-155.

Velleman, J. D. (2003). Don't Worry, Feel Guilty. *Royal Institute of Philosophy Supplement*, *52*, 235-248.

Walen, A. (2020). Retributive Justice. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/justice-retributive/

Wallace, R. H. (2019). The Tension in Critical Compatibilism. *Ethical Theory and Moral Practice*, *24*(1), 321-332. https://doi.org/10.1007/s10677-019-10038-2

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

Wallace, R. J. (2013). *The View from Here: On Affirmation, Attachment, and the Limits of Regret* Oxford University Press.

Waller, B. (2011). *Against Moral Responsibility*. MIT Press.

Watanabe, E., & Ohtsubo, Y. (2012). Costly apology and self-punishment after an unintentional transgression. *Journal of Evolutionary Psychology*, *10*(3), 87-105. https://doi.org/10.1556/JEP.10.2012.3.1

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

Watson, G. (2004a). Responsibility and the Limits of Evil: Variations of a Strawsonian Theme. In *Agency and Answerability* (pp. 219-259). Oxford University Press.

Watson, G. (2004b). Two Faces of Responsibility. In *Agency and Answerability: Selected Essays* (pp. 260-287). Oxford University Press.

Watson, G. (2011). The Trouble with Psychopaths. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (pp. 307-324). Oxford University Press.

Watson, G. (2013). Psychopathic Agency and Prudential Deficits. *Proceedings of the Aristotelian Society*, *113*(3), 269-292.

Williams, B. (1973a). Ethical Consistency. In *Problems of the Self* (pp. 166-186). Cambridge University Press.

Williams, B. (1973b). The Makropulos case: reflections on the tedium of immortality. In *Problems of the Self* (pp. 82-100). Cambrdige University Press.

Williams, B. (1976). Moral Luck. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *50*, 115-151.

Williams, B. (1981a). Internal and External Reasons. In *Moral Luck* (pp. 101-113). Cambridge University Press.

Williams, B. (1981b). Moral Luck. In *Moral Luck* (pp. 20-39). Cambridge University Press.

Williams, B. (1981c). Persons, Character, and Morality. In J. Rachels (Ed.), *Moral Luck: Philosophical Papers 1973–1980* (pp. 1-19). Cambridge University Press.

Williams, B. (1993). *Shame and Necessity*. University of California Press.

Williams, B. (1995a). How Free Does the Will Need to Be? In *Making Sense of Humanity* (pp. 3-21). Cambridge University Press.

Williams, B. (1995b). Internal Reasons and the Obscurity of Blame. In *Making Sense of Humanity* (pp. 35-45). Cambridge University Press.

Williams, B. (1995c). Moral Luck: A Postscript. In *Making Sense of Humanity* (pp. 241–247). Cambridge University Press.

Williams, B. (1999). Seminar with Bernard Williams. *Ethical Perspectives*, *6*(3-4), 243-265.
Williams, B. (2011). *Ethics and the Limits of Philosophy*. Routledge. (1985)
Williams, B., & Smart, J. J. C. (1973). *Utilitarianism For and Against*. Cambridge University Press.
Wojtowicz, J. (2022). The Purity of Agent-Regret. *Philosophy*, *97*(1), 71-90.
Wolf, S. (1982). Moral Saints. *Journal of Philosophy*, *79*(8), 419-439.
Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46-62). Cambridge University Press.
Wolf, S. (1994). *Freedom Within Reason*. Oxford University Press.
Wolf, S. (1997). Meaning and morality. *Proceedings of the Aristotelian Society*, *97*(3), 299–315.
Wolf, S. (2001). The Moral of Moral Luck. *Philosophic Exchange*, *31*(1), 4-19.
Woo, J. (2023). Moral Guilt without Blameworthiness. *Southwest Philosophy Review*, *39*(1), 201-208.
Zhao, M., & MacKenzie, J. (forthcoming). Survivor guilt. *Philosophical Studies*.

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

Responsibility and Ambivalence

I use the concept of *ambivalence*—the state of being faced with a choice that cannot be resolved without sacrificing something of value—to approach five contemporary debates in the philosophy of moral responsibility: (1) psychopathy, (2) free will, (3) the emotion of guilt, (4) regret and indirect moral luck, and (5) moral demandingness. Rather than arguing for one theory or another, acknowledging ambivalence paves the way for resolving these debates by reconciling the opposing sides.

LUND UNIVERSITY

Department of Philosophy
The Joint Faculties of Humanities and Theology
Lund University

ISBN 978-91-89415-92-8