Reconceptualising Blame, Blameworthiness and Basic Desert:

A New Perspective on Moral Responsibility (and Moral Responsibility Skepticism)

Thesis submitted to the University of Nottingham for the degree of

Doctor of Philosophy

September, 2025

Sean Shields

Student ID: 20272851

Supervisors: Chris Woodard, Zach Hoskins, Jussi Suikkanen

School of Humanities, Faculty of Arts

University of Nottingham

Acknowledgments

The philosophical beliefs I have nurtured in during my research for this thesis are supposed to make me skeptical about the worthiness of things like desert, praiseworthiness and gratitude. But I'm happy to throw all of that of the window when it comes to my principal supervisor Chris Woodard. He has been an inexhaustible source of patience, support, and kindness – never mind a fantastic person to be able to talk through all the philosophical stuff with. I feel very fortunate to have had you as my supervisor, Chris. Thank you.

Many thanks also to my secondary supervisors Jussi Suikkanen and Zach Hoskins, who have provided enormously valuable conversation and feedback, and been available whenever called upon. Thanks too to Aness Webster, the supervisor who I kicked everything off with, and who helped me to hone some of the initial ideas and instil my excitement about the project.

A big thanks to my PhD cohort at Nottingham, who I have enjoyed spending these years with. A particular mention to PhD Riana Popat who was a huge source of inspiration in her own PhD journey. Looking forward to finally being able to celebrate with you (we both made it in the end!).

Thank you to David Turner in the Disability Support Services team. Our conversations towards the end of the PhD were extremely helpful and have helped to give me a sense of things to take forward from the experience.

This thesis also owes a huge debt, of course, to those working in the field whose ideas I discuss in the thesis. I have been fortunate enough to meet some of you and benefitted from some very useful conversations. Shout out to the Lund Gothenburg Moral Responsibility Project and the PhD students there, who welcomed me when I spent three months at Lund University on a study visit.

I would also like to extend my thanks to my examiners, Anneli Jefferson and Joe Cunningham for their time in reading and examining my thesis.

Thank you to the Midlands4Cities DTP, and by extension, the AHRC, for funding my doctoral studies.

These last two thanks are big ones!

First to Gemma for being endlessly supportive and helpful, as well as a top-notch proofreader. It goes without saying I couldn't have made it without you. Sorry that it's taken me a bit longer than planned to get this out of the way – hopefully more time for all those exciting things we've talked about now! All my love.

Finally, my biggest debt, as always in my life, is to my parents. During what has been not the easiest of times, you have remained a constant source of support for me, as you have throughout all my other endeavours. I'm looking forward to spending more time together unburdened by thoughts of the PhD! I dedicate this thesis to you.

Table of Contents

Introduction

This thesis makes two broad sets of arguments, as reflected in its two distinct parts. In Part One, the first set of arguments concern how we might profitably reconceptualise moral responsibility. What motivates this is partly a sense of frustration at how concepts which are undoubtedly central to the moral responsibility debate – blame, blameworthiness, and moral responsibility itself – have been so variably, and quite often ambiguously, interpreted. As a result, it can be difficult to know to what extent (apparently) competing theories of moral responsibility are in fact talking about the same thing.

Of course, in any area of philosophical enquiry, assumptions about basic concepts must be made. There is a danger of spending too long defining our terms and never reaching the more interesting destinations of enquiry to which we might want to set sail. But my growing impression was that the moral responsibility debate is *especially* afflicted with a lack of clarity about some of its most fundamental concepts. I felt there was room, and need, for a new perspective. In the first half of the thesis, I therefore propose an original *overarching conception of moral responsibility*. If theories of moral responsibility are in the same business, this overarching conception represents my attempt at defining the business they are in.

The second set of arguments, in Part Two, relate to a specific kind of moral responsibility: *basic desert*. During my initial research into the moral responsibility literature, I became convinced that this is a species of moral responsibility with especially profound importance. In particular, I found the established arguments in favour of *skepticism* about basic desert enormously powerful, and I was interested in exploring their practical significance. However, as my research developed, I increasingly sensed that the notion of basic desert had not been as clearly articulated as I had initially thought. The question of what basic desert *is* remains unsettled, and this needs attention first. Part Two of the thesis therefore addresses this, and aims to establish the kind of moral responsibility which the established arguments for basic desert skepticism in fact serve to undermine. The central upshot of my discussion is that what has been presented as basic desert skepticism can in fact be reconstrued as basic *reason* skepticism. This recentres the skepticism in what I think is a clarifying way: it now principally reveals how familiar features of our agency which are typically taken to have an important kind of normative significance, in fact *lack* such significance. The practical implications of such skepticism are thereby potentially even more wide-ranging than has been previously appreciated.

The two parts of the thesis therefore address distinct issues, but they share a similar motivation. They are each a response to what I perceive as a lack of clarity in how we ought to understand the relevant concepts. The conceptual framework I set out in Part One of the thesis, I hope to show, provides a clarifying lens for the arguments made in Part Two. Nonetheless the arguments of Part Two do not depend upon on accepting the arguments of Part One. In this sense the two parts of the thesis therefore stand apart and have independent merit.

Chapter Plan

Both parts of the thesis are divided into two chapters. This, in the end, seemed like an intuitive way of breaking up the material, although the resulting four chapters are each rather hefty. Here is a brief guide to what follows in the chapters.

Chapter 1

I open the thesis by looking to get an initial grip on the notion of moral responsibility, first by contrasting it with two alternative kinds of responsibility (causal and assigned). I then begin to construct my overarching conception of moral responsibility by setting out the idea of *agential features*, drawing from examples which we find in various theories of moral responsibility. I propose that these features, which bestow agents with control over their actions and attitudes, are an indispensable component of moral responsibility. I then argue that agential features are likewise essential for blameworthiness – but they must be suitably deployed in a negative moral performance. I contrast these essential constituents of blameworthiness with facts which bear upon blame's justification, but not blameworthiness.

Chapter 2

I continue to build my overarching conception of moral responsibility by setting out a theory of blame. I first survey some prominent examples of blame found in the literature. These inform my *Appraisal Theory* of blame. This theory sets a single qualifying requirement for blame responses: an appraisal of the essential constituents of blameworthiness. It therefore amounts to an extremely thin and undemanding theory of blame. I discuss some implications of the view and defend it against some possible objections. I endorse an ambitious version of the *Appraisal Theory*, whilst also proposing a more moderate version. Either way, the *Appraisal Theory* complements the conception of agential features and blameworthiness I offered in Chapter 1, and it completes my overarching conception of moral responsibility. This complete conception generates a formula for how theories of moral responsibility are constructed. I end by setting up the problem for Part Two, which aims to fill out this formula for the theory of moral responsibility which is targeted by basic desert skepticism. There are gaps in how basic desert has so far been specified – it clearly implicates what I call *sheer agential features*, but it is unclear what kind of blame and blameworthiness is at stake.

Chapter 3

Here I interrogate more closely how the basic desert skeptics – principally Derk Pereboom and Gregg Caruso – have conceived basic desert moral responsibility. I attribute to them a position called *Retributivism*. I then set out a prominent alternative account of basic desert offered by Leo Menges, the *Claim Forfeiture* theory. Menges rejects the commitments of *Retributivism* as viable for a basic desert thesis. In response, I defend *Retributivism* against these objections, arguing that the desiderata which Menges refers to in making them are not fit for purpose. Contra Menges, both the *Claim Forfeiture* and *Retributivism* theories are still in the running as the correct interpretation of the kind of moral responsibility at stake in basic desert skepticism. I then set out a third possible candidate for the correct understanding of *Basic Reason* moral responsibility. The natural methodology for adjudicating which of these theories – if any – is the theory of moral responsibility targeted by basic desert skepticism is to turn to the skeptical considerations themselves.

Chapter 4

This chapter takes up the work promised in Chapter 3. I first summarise Pereboom's manipulation argument, before raising conceptual concerns about how Pereboom has framed the argument, and which work to undermine its effectiveness. We can address these issues by reframing the manipulation argument, so that it tests for intuitions about moral responsibility in the three senses set out in Chapter 3. I argue that the manipulation argument is in fact effective in undermining *each* of these kinds of moral responsibility (*Retributivism*, *Claim Forfeiture* and *Basic Reason*). These findings cast the *Retributivism* and *Claim Forfeiture* theories in a new light, and they also reveal the manipulation argument's true potency. The upshot is that we can reconstrue basic desert skepticism (as it has been known) in conceptually more straightforward terms. Such skepticism establishes a powerful conclusion: an important set of our agential features lack the freestanding normative significance which they are very often credited with by theorists of moral responsibility and the folk alike.

PART ONE: CONCEPTUALISING MORAL RESPONSIBILITY

Chapter 1 - Agential Features & Blameworthiness

1.1. Introduction

In Part One of this thesis, I aim to construct an overarching conception of moral responsibility. This chapter begins this work, by setting out the first two of the three components which comprise this overarching conception. First, agential features. Second, blameworthiness and praiseworthiness. (As I will indicate, as my discussion proceeds my focus will be predominantly on blameworthiness and blame, and only derivatively on praiseworthiness). In the course of setting out these two components, I will survey some leading theories of moral responsibility. This principally serves to establish the serious divergence of ways in which moral responsibility – and its attendant concepts which include agential control and blameworthiness – has been understood. My chief aim in the first half of the thesis is to try and impose some conceptual order on this often-confusing state of affairs.

Having discussed and clarified these first two components, I will then be in position – in Chapter 2 – to set out the third: a theory of blame (and praise). This will complete my overarching conception of moral responsibility. One of the functions of Section 1.2 is to give a fuller introduction to the considerations which motivate this chapter, where it sits within the broader thesis, and what it aims to achieve. So, with this in mind I will jump straight into a breakdown of the chapter.

This chapter has three main sections and proceeds as follows. In Section 1.2 I begin the task of zeroing in on what moral responsibility is. First, by contrasting it with two kinds of responsibility which are more straightforwardly and uncontroversially identified: causal and assigned. The contrast I especially seek to draw is how moral responsibility distinctively brings with it an inherent response-worthiness. I then register the wide diversity of terms in which theories of moral responsibility have been expressed, and give an overview of my overarching conception of moral responsibility which aims to make some sense of this diversity. In Section 1.3 I set out the first component of this conception: *agential features*. These are those features of a person which confer upon them control – very expansively interpreted – over their actions, attitudes or character. Theories of moral responsibility invariably identify such a feature and afford it a privileged and irreducible role in explaining moral responsibility. I also note how those agential features highlighted by these theories are in a sense *humdrum,* insofar as they are typically composed out of garden-variety mental states or events. I conclude this section with a brief detour into instrumentalist theories of moral responsibility. I suggest these do not in fact qualify as theories of moral responsibility (at least as I am conceiving of it) given the terms I have set. They fail to afford agential features the right kind of role in explaining moral responsibility (and blameworthiness). In Section 1.4, I turn to consider the second component of my overarching conception of moral responsibility: blameworthiness and praiseworthiness. I suggest that we can think of these as statuses which generate normative reasons to blame and praise. This much is fairly commonsensical. But I argue that these are reasons which, at minimum, are grounded in agential features when suitably deployed in non-neutral moral performance. Deployed agential features in non-neutral performance therefore comprise the essential constituents of blameworthiness (and praiseworthiness). I then make a distinction between *control-conferring* and epistemic agential features. I conclude the chapter by identifying some facts which are *not* part of the essential constituents of blameworthiness, but which do often influence the all-things-considered justifiability of blaming.

1.2. The Relevant Sense of Responsibility

We can begin by distinguishing three distinct senses of responsibility – different ways in which entities can be said to be responsible. The first two – causal responsibility and assigned responsibility – are expressly *not* the focus of this thesis and can quickly be set aside. But there is still some value in drawing a contrast between these two senses of responsibility and the third sense – *moral* responsibility – which is my exclusive concern. This will give some initial shape to this third kind of responsibility. Given that the term 'responsibility' is often used in all three senses, by philosophers and the folk alike, it is also important to clearly distinguish them at the outset to allow for maximum clarity as we proceed.

1.2.1. Two Uncontroversial Kinds of Responsibility: Causal and Assigned

First, there is *causal* responsibility. Put simply, this is the familiar idea that one occurrence or action is in a causal relationship with another. We can say X is causally responsible for Y wherever X has a causal impact on Y, thus helping to bring Y about. If X is but one of several causes of Y, we might say X is *partially* causally responsible for Y, in the sense that other entities than X share causal responsibility for Y.

People are very evidently capable of causal responsibility (the movement of my fingers is causally responsible for the creation of this text; my decision to type is causally responsible for both of those things). But clearly this kind of responsibility is not limited only to people. Countless other kinds of things are responsible in this sense. Each atom is causally responsible (at least partially) for the movement of its near neighbours. A tsunami is causally responsible for devastating a coastal village; an anaconda for striking fear into a passing hiker; an image glimpsed as a child for a memory experienced sixty years later.

Notably, that something possesses causal responsibility does not by itself seem to tell us anything about whether any particular response, either towards the responsible entity or anything else, is warranted or appropriate or fitting. Other facts need to be present if causal responsibility is to gain such normative significance. For example, the mere fact that a football player is causally responsible for a football ending up in the opposition's goal by itself does not make any particular response worthy. Particular responses will only be worthy if other facts obtain. In this case, for example, facts about which football team you support. Joy would become fitting *if* you support the team of the player causally responsible for the ball going into the goal, as well as love and gratitude towards the player. If your loyalties lie with the opposition team, then despair (and perhaps opprobrium towards the responsible player) would instead be apt. But without such additional facts, causal responsibility does not generate any kind of normative consideration in favour of any particular response. Indeed, the huge majority of the vast number of instances of causal responsibility pass off without any call for a response. This lack of an inherent *response-worthiness*, as we might call it, marks an important contrast with moral responsibility, as I will highlight in Section 1.4.

A second kind of responsibility is what I will term *assigned* responsibility.[1] This instead concerns having responsibility in the sense of having a duty. In virtue of being

[1] For a classic discussion of these first two kinds of responsibility, causal and assigned, see Hart 1968: 210-237. Hart refers to assigned responsibility as *role-responsibility*.

assigned (or assigning ourselves) to a particular role, we acquire duties (or *responsibilities*) to ensure certain actions are taken within a particular domain of activity, and that these actions are performed according to certain expectations or standards.[2] Assigned responsibilities therefore always have a prospective dimension, namely the duties associated with the role which are to be carried out. For example, a newly appointed Home Secretary gains (assigned) responsibility for implementing policy in areas such as law enforcement and immigration; a PhD student has a responsibility to complete their thesis. Each is (prospectively) responsible for these things simply in virtue of occupying their respective role. They need not yet have taken any actions within the relevant domain of activity – they need not be causally responsible for anything – to gain the status of 'responsible' in this assigned sense. The assignation alone – the occupying of a particular role – is sufficient to render them responsible.

In addition, assigned responsibilities typically also have a retrospective dimension, concerning how successfully the duties associated with the role are carried out. If these duties are not performed well, burdens are often incurred. One can be liable to blame and sanction; you might also acquire the duty to take reparative actions (this might be thought of as a new assigned responsibility conferred by the role under those particular circumstances). Whereas if you perform the duties successfully, you are instead usually entitled to certain benefits, including praise. If the Home Secretary implements a successful immigration policy, say, she might be rewarded with praise and the prospect of promotion. But if the policy fails disastrously then it will be her, in virtue of her assigned responsibility, who becomes the appropriate target of blame. She will be tasked with fixing things and may face penalties, perhaps even removal from office, due to her failure to fulfil her responsibilities.

[2] The scope of *actions* here is wide. It is meant to include omissions as well as commissions, and also the consequences of both. These are each things for which responsibility is typically assigned. (A steward at a football match is assigned responsibility to help oversee the safety of the crowd, and fulfilling this duty requires that they *refrain* from observing the game. He is assigned responsibility not merely to engage in such an omissive act, but also to ensure it has the desired *consequence* of upholding safety). Although, in principle, it would seem there is no limit to what we can assign responsibility for, provided one has the authority to assign duties.

Unlike causal responsibility, then, assigned responsibility – specifically with regards to its retrospective dimension – typically does bring with it response-worthiness.[3] But the exact nature of the responses it can render apt will depend on the particular specification of the role and responsibilities in question, and could vary widely. Home Secretaries and PhD students are subject to very different kinds of response in light of how they carry out their roles. This represents a point of difference with moral responsibility where, I aim to show, we can more tightly specify the nature of the response which such responsibility renders worthy.

It seems that any object whatsoever could, at least in principle, possess assigned responsibility. All that is needed to acquire it is an act of assignation: it is *possible*, if misguided, to assign the responsibilities of Home Secretary to a potato or to a giraffe. In practice, however, we tend to assign responsibilities to those we expect to be capable of fulfilling the duties of the role. As such, and unlike causal responsibility, assigned responsibility is mostly a status reserved for people.[4] Whereas unlike both causal and assigned responsibility, moral responsibility is a status which can be held by only a restricted set of things. As I will argue in Section 1.3, this set consists of entities who possess what I call agential features.

Finally, assigned and causal responsibility are two forms of responsibility which, I take it, uncontroversially exist. However, there remains one further sense of responsibility – a different sense in which we might ask 'Is the Home Secretary responsible for this policy and its effects?'.[5] Are they *morally* responsible? The existence of moral responsibility – at least certain varieties of it – has been hotly contested. This is the kind of moral responsibility in which I am interested.

---

[3] I say *typically* because it seems at least conceivable that one might be assigned responsibility without any associated expectations about how responses (burdens or benefits) will become appropriate or worthy in light of how well the duties are carried out. (Suppose the assigner is a disillusioned manager, entirely and obviously uninterested in how their employee performs in the role). But in practice we nearly always see that assigned responsibilities have the retrospective dimension – a recognition that how one performs in the role will generate normative considerations in favour of certain responses.

[4] Albeit not exclusively, with due respect to (inter alia) guide dogs and carrier pigeons.

[5] I should say there is *at least* one additional sense of responsibility. Perhaps a complete taxonomy of responsibility will contain more than just these three varieties (causal, assigned, and moral). But dividing up the terrain in this way, so as to mark out moral responsibility from these other two kinds, is adequate for my purposes. This tri-partite division also accords with most writing on the topic – it is fairly common to distinguish between these three kinds of responsibility at the outset of any treatment of moral responsibility. For instance, see McKenna & Pereboom (2016: 12) or Talbert (2024).

1.2.2. Responsibility of the Third (and Controversial) Kind: Moral Responsibility

What, then, is moral responsibility? Whilst it is the aim of these next two chapters to give a much fuller answer to this question, we can begin with a very rough approximation. I think we can reasonably say it is a kind of responsibility which we are understood to possess in light of our moral status or performance. It is usually understood to also require possession of particular capacities – most notably some kind of freedom or control. Depending on how the relevant capacities are exercised, and in connection with our moral status or performance, we become liable or worthy for particular kinds of response. This is what I mean by the *response-worthiness* which is distinctively inherent to moral responsibility. This response-worthiness is most often expressed in term of our blameworthiness and praiseworthiness. It is responses of blame and praise of which the morally responsible agent is distinctively worthy.

These will all be broadly familiar ideas and concepts. But, despite their familiarity, these concepts (and their interrelationship) can be, and have been, interpreted in wildly different ways. As such, this rough approximation is not especially informative – it amounts to only a vague indication of what we might take moral responsibility to mean. I aim, in presenting my own conception of moral responsibility, to put meat on these very bare bones.

Testament to how the relevant concepts have been so widely interpreted is the large, diverse and growing number of views advanced as theories of moral responsibility. I will not attempt any especially novel taxonomy. But a rough categorisation of these theories, along with some notable proponents of each, might look something like this:[6]

(a) consequentialist theories or instrumentalist theories (e.g., Smart 1961, Vargas 2013; Jefferson 2019);

(b) agent-causal theories (e.g., Chisholm 1976; O'Connor 2009a);

(c) event-causal theories (e.g., Balaguer 2010; Kane 1996);

[6] This taxonomy is adapted from one offered by Shoemaker, referring to its complexity as "alphabet soup, a mess" (2020: 208). Dividing things up in this way is inevitably somewhat artificial; it may be possible to integrate more than one of these views into a singular theory of moral responsibility. For example, as we shall see, some more recent consequentialist theories of moral responsibility also incorporate the conceptual resources of reasons-responsiveness.

(d) theories of "ultimate" or "absolute" moral responsibility (e.g., Strawson 1994);

(e) ledger theories (e.g., Zimmerman 1988; Haji 1998);

(f) quality of will theories (e.g., Strawson 1963; Arpaly 2006; Shoemaker 2013, 2015; Arpaly and Schroeder 2014);

(g) conversational theories (e.g., McKenna 2012);

(h) expressive or communicative theories (e.g., Macnamara 2015);

(i) judgment-sensitive theories (e.g., Scanlon 1998; Smith 2005);

(j) reasons-responsive theories (e.g., Fischer and Ravizza 1998; Brink and Nelkin 2013; Sartorio 2016);

(k) rational abilities theories (e.g., Wolf 1990; Nelkin 2011);

(l) fair avoidability theories (e.g., Brink and Nelkin 2013).

Without claiming this list to be in any way exhaustive, this already amounts to a bewildering array of theories.[7] This proliferation of moral responsibility scholarship has brought an ever-expanding lexicon and more fine-grained distinctions – what Knobe and Doris call a "swamp of technicalities" (2010: 324). As a result, it can be increasingly difficult to discern precisely what these theories have in common. In wading into this swamp my goal is not to elucidate, much less defend, any particular theory of moral responsibility. Rather it is to identify some essential ingredients which these theories, despite their diversity, each share. Taking these ingredients together, I will establish a unified and overarching *conception* of what moral responsibility is.[8] The relevant ingredients can be specified in very different ways, which is what gives rise to distinctive theories of moral responsibility. But I propose that each of these

[7] We ought also to acknowledge the significant cultural variation in beliefs about moral responsibility, across time and geography, many of which may not fit neatly under the rubric of any of these (largely Western) philosophical theories. Tamler Sommers (2012) offers a comprehensive survey of this cultural diversity.

[8] I appreciate this may deviate slightly from the traditional usage of *conception*. I use this word to reflect how I do not defend any particular theory of moral responsibility. Rather, I am proposing how the concept of moral responsibility might be conceived in an original and illuminating way, a conception which makes good sense of the various theories of moral responsibility we encounter in the literature. So, different theories may (and I argue in most cases do) comply with this same conception.

candidate theories of moral responsibility, in virtue of their common conceptual ingredients, will comply with the overarching conception.

Here is a preview of this overarching conception of moral responsibility which I will elaborate across the following two chapters. This conception analyses moral responsibility in terms of three distinct but interdependent components (the 'ingredients'). First, *agential features*. This is the term I give to those features of agents which endow them with an important kind of control that is necessary for moral responsibility. I will set out a very inclusive understanding – what I call the 'skeleton concept' – of agential features. The second component is *blameworthiness* and *praiseworthiness*. These I understand as statuses acquired by agents where, at minimum, there exists a suitable relationship between their agential features and their actions or attitudes. This relationship generates a normative reason in favour of responding to them with praise or blame. I give my proposal for the essential constituents of blameworthiness and praiseworthiness in Section 1.4. The third component is the *blame* and *praise* response itself. My discussion of this component takes place in Chapter 2, where I recommend a way of conceptualising these distinctive kinds of response.

Clearly, this conceptual terrain is not new (it echoes the rough approximation of moral responsibility with which I began). Perhaps few would object that the three conceptual ingredients I identify are interdependent, and integral to any conception of moral responsibility. What is much less clear, however, is how we ought to understand these concepts. As we shall see, competing theories of moral responsibility frequently deploy concepts such as blame, blameworthiness, and moral responsibility itself, to denote markedly different things. Quite often it is simply unclear what they denote.[9] (We will see this, for instance, when I assess how skeptics have described the blame and praise which is at stake in basic desert moral responsibility). I therefore hope to

[9] I am by no means the first to make such observations about the state of the moral responsibility literature. Warmke, for example, puts it in terms which I wholeheartedly endorse: "[The] philosophical literature on moral responsibility and cognate topics over the past fifty years or so is a messy and at times confused literature. The primary culprit for this messiness and confusion is that philosophers have often not been careful to explain clearly what they have in mind when they discuss 'moral responsibility'. The result has been that a lot of philosophers have spent a lot of time talking past each other. It has only been fairly recently, however, that philosophers have begun to untangle and delineate all of the various ways of understanding 'moral responsibility'" (2011: 198). These first two chapters are effectively my contribution to this campaign of disentanglement and delineation.

bring some needed clarity to things by offering a fresh interpretation of these familiar concepts, culminating in a conception of moral responsibility which I believe can accommodate and make good sense of the existing theories. This clarificatory exercise will also set things up for my analysis of one species of theory in particular – basic desert moral responsibility – which becomes my task in the second half of the thesis.

1.3. Agential Features

1.3.1. Some Examples

The first item which I claim is common to all theories of moral responsibility – with one notable exception, which I will discuss – is *agential features*. Consistently, such theories will identify a feature of this kind, and it will play a fundamental role in their respective accounts of moral responsibility. To delineate the concept of an agential feature, I will first table some prominent examples of what I have in mind, taken from established theories of moral responsibility (including some of those mentioned in the previous section). Then, drawing out what is common to each of these examples, I will derive what I call the *skeleton concept* of an agential feature.

A good place to begin is with P.F. Strawson, and his famous essay *Freedom and Resentment*. The basis of Strawson's account is empirical: he observes the occasions when we treat each other (including ourselves) as morally responsible. In particular, where we are prone to respond with what Strawson called the "reactive attitudes", which are emotions such as resentment, indignation, guilt, and (more positively) gratitude (1963: 192-199). He observes how we tend to experience the reactive attitudes in response "to the quality of others' wills" towards ourselves or third parties. From this he concludes that we are morally responsible for our actions and attitudes wherever these reflect the level of "goodwill or regard" (or lack thereof) which we have towards others (1963: 199-200). For Strawson, then, exhibiting this quality of *regard* is a necessary condition for being morally responsible.

Strawson's account is a hugely influential precursor to many 'quality of will' theories. Some of these theories understand moral responsibility with reference to a different sort of quality which our will can exhibit, namely that of our *character* (rather than quality of regard). These are often described as 'Real Self' theories – they understand moral responsibility in terms of a suitable correspondence between our real self (or character) and our attitudes or actions. Harry Frankfurt offered a *desires*-based version of such an account. According to his view, it is our "second-order volitions" – second-order desires about which first-order desires we wish to motivate our actions – which explain how we can be morally responsible (1971: 10). Frankfurt argues that we are morally responsible for those actions which are in accordance with these second-order desires. Frankfurt (1988, 1992) subsequently refined this initial view, so that in addition to having the requisite hierarchy of desires we must also *identify* with our motivating desires, if we are to be morally responsible for their effects. Gary Watson (1975) offered a different Real Self theory, instead arguing that the attitudes and actions for which we are morally responsible are those which cohere with our values, rather than our higher-order desires. Real Self views, such as Watson's and Frankfurt's, have a distinctively hierarchical or mesh structure. They maintain that to be morally responsible for actions or attitudes, our values or desires must be configured or meshed such that we – or more precisely our real self or our character – *endorse* those actions or attitudes. Such views are also sometimes referred to as "attributionist", since moral responsibility requires being able to attribute our actions or attitudes to our real selves (Talbert 2024).[10]

A popular family of theories says that moral responsibility requires possession of a different capacity: reasons-responsiveness. A classic statement of such a view comes from Fischer and Ravizza (1998). They articulate reasons-responsiveness in terms of "guidance control," whereby an agent guides their behaviour via sensitivity to the rational considerations which bear upon their actions (1998: 29–34). There is a diverse range of reasons-responsiveness views.[11] What unites them is the idea that moral responsibility requires this ability to recognise and react to the pertinent moral reasons. Some view reasons-responsiveness as a further, third kind of quality which our wills can exhibit (Shoemaker 2015: 64-86). That is, a quality of judgment about reasons (as opposed to quality of regard or character). On views such as these, then, it is the

[10] The notion of a *Real Self* is usually credited to Susan Wolf (1990). More recent prominent Real Self (or Deep Self or attributionist) theories include Arpaly and Schroeder (2014); Doris (2015); Sripada (2016). Shoemaker (2015) provides a very useful overview of quality of will approaches more generally.

[11] Other prominent reasons-responsiveness theories include Brink and Nelkin (2013), McKenna (2013), Nelkin (2011), Sartorio (2016), and Wolf (1990).

capacity for reasons-responsiveness, or to exhibit quality of judgment, which is a necessary condition for moral responsibility.

We might set a lower bar for moral responsibility. Another possible theory says that we are morally responsible for our actions simply where they flow intentionally from our desires, without external impediments. Thomas Hobbes was a notable early proponent of this sort of idea. In his *Leviathan* he writes that that a person is free wherever he finds "no stop, in doing what he has the will, desire, or inclination to doe" (1651: 108).[12] Here what is specified is a straightforward form of *voluntariness*: perhaps it is enough for moral responsibility if we do what we intend to do. This amounts to possessing a kind of conditional freedom – we are free in the sense that had we intended otherwise, we would have done otherwise.

Each of the theories canvassed so far are compatibilist proposals – the conditions they place on being morally responsible are compatible with determinism.[13] A different branch of theories – libertarian views – are incompatibilist. These theories insist that moral responsibility and determinism do not mix – there must be some degree of *indeterminacy* in the process leading up to a decision or action if we are to be morally responsible for them. There are two dominant schools of libertarianism. The first – event-causal libertarianism – accepts that our actions are deterministically caused by prior events (such as our beliefs or desires). But it insists that these prior events, at least sometimes, are themselves indeterminate – they are not necessitated by deterministic forces to occur – and this creates the space for moral responsibility. Robert Kane (1996, 2007) has developed perhaps the most well-known event-causal theory. On his view, our character determines what we do. But our character is the product of '*self-forming actions*', which are "choices or actions voluntarily performed in the past for having the character and motives he or she now has" (2007: 14). These are the choices and actions we take "when we are torn between competing visions of

[12] Hume (1739) and Ayer (1954), amongst others, also advocated a view along these lines, although relatively few today subscribe to conditional freedom as sufficient for moral responsibility. Indeed, we should be careful to note that both Hobbes and Hume were themselves describing what they saw as the condition for *freedom of action*, rather than explicitly one which bears upon moral responsibility. As I discuss at the end of this section, my aim for now is just to adduce some examples of what I will identify as *agential features*. I am less concerned whether the authors who describe them themselves take them to be important for moral responsibility, or something else.

[13] Determinism standardly being understood to mean a world in which the past, in combination with the laws of nature, necessitates a single unique future (Mele 2006: 3; van Inwagen 1983: 3).

what we should do or become" (2007: 26). Kane maintains that these self-forming choices are *not* deterministically caused – and this provides the requisite indeterminacy needed for moral responsibility. The second dominant school of libertarianism – agent-causal libertarianism – locates the requisite indeterminism elsewhere. Here, rather than a prior event, it is the agent *themselves* acting as an irreducible substance which has the power to directly cause decisions and actions without itself being causally determined. As notable agent-causalist Timothy O'Connor puts it, agent substances have the power to non-deterministically cause "the coming to be of an executive intention (an intention to act here and now)", and it is the exercise of this power which renders one morally responsible (2009b: 119). Of course, these are all intricate positions which warrant much longer explication and discussion. But crucially, across these libertarian views, it is the indeterministic causation of our choices which purportedly explains why we are morally responsible for them.[14]

1.3.2. The Skeleton Concept of Agential Features

This provides merely a snapshot of some leading theories of moral responsibility. This snapshot in no way represents the richness and complexity of these theoretical approaches, nor the diversity of views which each can accommodate. It is intended only to highlight a singular feature which these theories all incorporate into their account and deem necessary for moral responsibility. The following are each different expressions of this *agential* feature: voluntariness; quality of will (whether understood in terms of regard, character, or judgment); reasons-responsiveness; higher-order endorsement or identification (via desires or values); libertarian freedom (whether indeterministic event-causation, or indeterministic agent-causation).

What do these things have in common? In my view they can each be understood as features of an agent which endow them with some measure of *control* over what they do or think. They are means by which we exert influence as *agents*, rather than being merely passive conduits of our behaviours or attitudes. These features are aptly

[14] Other prominent event-causal accounts include Balaguer 2010; Ekstrom 2000; Franklin 2011; Kane 1996). Prominent agent-causal theories include Chisholm 1976; Clarke 1993, 2003; Griffith 2010; Kant 1781; O'Connor 2009a; Taylor 1974.

described as *agential,* then, because to possess control over our actions is to be afforded some degree of agency over them.[15]

We arrive then at what I will call the *skeleton concept* of an agential feature. This denotes any feature of an agent which allows them to wield control over what they do or who they are.[16] This skeleton concept is very expansive in the sense that it might be specified in very different ways. The theories of moral responsibility I have canvassed each add flesh to the skeleton concept by specifying a *particular* agential feature which they deem to be requisite for moral responsibility. Whereas the skeleton concept is stripped of these more specific commitments, and simply denotes *any* control-conferring feature.[17]

Although I have considered only a sample of theories, my observation is that invariably theories of moral responsibility will similarly identify a specific agential feature (or occasionally a combination thereof). In each theory, this feature – and the control it confers – play an essential and irreducible role in their explanation of how moral responsibility arises. This finding suggests we should understand agential features to be an indispensable element of that phenomenon – moral responsibility –

[15] It is worth briefly mentioning here an objection that has been levelled at event-causal libertarian accounts. Recall, these accounts cite indeterminate events in the formation of our beliefs and desires, which constitute our character, as being essential for moral responsibility. The worry is that, given the indeterminacy, it is unclear in what sense it is the *agent* who meaningfully settles which belief or desires are formed (or, therefore, the subsequent decisions which issue from their character). This criticism of the "disappearing agent" has been most strongly pressed by Pereboom (2014: 32-49). It is a worry about whether the agent can be said to exercise any kind of control which might allow for moral responsibility. In my terms, then, the objection throws into question whether any agential feature is in fact even implicated in event-causal libertarian views. Still, event-causal libertarians have insisted that such agents *do* gain a crucial kind of control. Kane for instance, describes the agent as exercising *rational or reflective self-control* when they perform the relevant self-forming actions. Even though these choices and actions are undetermined, they are still voluntary, intentional and made without coercion (1996: 179). So, we can read Kane and likeminded event-causal libertarians as asserting the presence of control-conferring agential features where skeptics like Pereboom deny it. I am therefore willing to include the indeterministic phenomena described in event-causal libertarian accounts amongst my list of agential features, provisional on how this dispute plays out.

[16] One important kind of feature which I have not yet mentioned are our *epistemic* capacities: those which equip us with knowledge in relation to our choices and actions. These capacities clearly endow agents with an important kind of control. Given the expansiveness of the skeleton concept, they plausibly ought to be considered as another kind of agential feature. I discuss this in Section 1.4.5.

[17] I do not claim the agential features I have enumerated here in any way exhaust those we find in theories of moral responsibility. I have picked out just a selection of prominent examples, and provided only a sketch of what they entail. But I believe the idea is rendered sufficiently clear for my purposes. In particular, it is not necessary to compile an exhaustive list of agential features in order to give an analysis of basic desert skepticism, which is my subsequent task. We will see that having a grasp of the skeleton concept is sufficient.

which I am seeking to get a handle on. Agential features therefore comprise the first component in my overarching conception of moral responsibility.

Before moving on from these examples, we should note that there is no *exclusive* relationship between agential features and moral responsibility. The majority of the theories I have mentioned so far are proposed as theories of moral responsibility. But not all: some implicate (what I articulate as) agential features, without being explicitly or principally advanced as theories about moral responsibility. For instance, Frankfurt presents his hierarchical desires-based theory as describing conditions for when "the person did what did freely and of his own free will" (1971: 19). He is, though, explicit that acting of our own free will in this way is a necessary condition of moral responsibility. (Whereas he deems *having* free will – as opposed to acting of one's own free will – to be something distinct, with different satisfaction conditions, and *not* required for moral responsibility). As noted above, Hobbes and Hume were theorising about freedom of action rather than moral responsibility, or free will.[18]

Indeed, many notable accounts of moral responsibility are simultaneously advanced as accounts of free will. This reflects how free will is often understood to represent the *control* condition on moral responsibility (e.g., Fischer and Ravizza 1998; Kane 1996; McKenna 2013; Mele 2006; Pereboom 2014; Sartorio 2016). If we take this view, then the agential features (in my sense) requisite for moral responsibility will overlap very closely with the capacity of free will. Occasionally, though, authors have instead interpreted free will in a way which is independent of considerations of moral responsibility (e.g., Clarke 2003; Ginet 1990; Vihvelin 2013). Bruce Waller is distinctive in being a skeptic about moral responsibility but not free will (2011: 43-58).[19]

I will remain agnostic about how to understand free will, and its relationship with moral responsibility. In fact, I see it as a clarifying aspect of my conception of moral responsibility that it avoids needing to take a stand on the intractable issue of free will. Again, how far individual accounts take themselves to be setting conditions on moral

---

[18] Indeed, Hume and Hobbes were amongst those who dismissed the notion of having freedom with respect to our will as incoherent (as discussed in McKenna & Pereboom 2016: 50-51).

[19] For additional complexity, some theorists argue that satisfying the conditions for moral responsibility qualifies us for other statuses too. For example, the status of personhood (Frankfurt 1971), or membership of the "moral community" (Stern 1974; Strawson 1963: 207; Shoemaker 2007).

responsibility, free will, or something else again is not my primary concern. (Indeed, even for those theories which *do* explicitly purport to describe conditions of moral responsibility, we cannot be sure without much closer scrutiny that they understand moral responsibility in a like manner). My aim here is simply to get onto the table some examples of what I identify to be *agential features*. What matters is that the theories in question each identify a feature of this sort. Therefore, whatever they profess to be theories about, they each exhibit what I deem to be the first essential component in any theory of moral responsibility.

1.3.3. A Looser Notion of Control

The agential features I listed in the previous section admittedly comprise a motley group. (Having the ability to recognise and respond to reasons, for instance, is a quite different thing to acting or forming attitudes in ways which express our real self. Both are quite different again from the exercise of libertarian freedom). To interpret all of these as forms of control requires a perhaps generously broad understanding of what control is. But I suggest they each describe a different way of rendering our actions or attitudes as things which we, as *agents*, in some sense make happen. To say they are different ways of possessing control is, I suggest, a reasonable way of capturing this idea.

The claim that possessing control is necessary for moral responsibility is not new. A control condition on moral responsibility is widely accepted, whereby a morally responsible agent "must *control* his behaviour in a suitable sense, in order to be morally responsible for it" (Fischer and Ravizza 1998: 13; original italics). However, the relevant control has most often been understood in terms of exercising *voluntary* control over intentional actions (McKenna 2008; Smith 2008). This narrower, voluntarist conception of control cannot accommodate the full range of agential features which I have adduced, such as exhibiting quality of regard or the higher-order endorsement of our actions.[20] These do not seem like operations of voluntary control. I am therefore advocating we use a broader notion of control, one which has

[20] The term *voluntarist* is taken from McKenna (2012: 188)

application to the full range of agential features which we find implicated across theories of moral responsibility (and beyond).[21]

If we understand agential features in the expansive sense which I recommend, then it implies that *every* intentional action necessarily involves at least one agential feature (voluntariness). Indeed, plausibly very many of our typical actions will exhibit more than one agential feature simultaneously. For instance, suppose that Julia begins to send money to a charity. Her action is the product of her voluntary intention to send the donation. Suppose it is also true that Julia desires to send the money, and that she has a second-order volition that this desire be effective in motivating her to act. Her donation also expresses the value which she places on the charity's work. It simultaneously is a response to her recognition of the various good moral reasons which favour giving to the charity. Each of these amounts to a distinct agential feature – a different sense in which Julia wields control with respect to her charitable donation. So, agents can often exhibit a bundle of agential features at once. It is noticeable however that theories of moral responsibility each tend to fix upon only *one* agential feature, privileging it as delivering the kind of control which is uniquely essential for explaining moral responsibility (hence why we can categorise these theories in terms of their associated agential feature).[22] In this sense these theories offer different but individually homogenous, 'agential feature-specific' accounts of what moral responsibility entails.

A final, important distinction to register relating to agential features concerns the *species* to which the particular kind of control which they describe belongs. Here there is a well-established bifurcation in the literature. First, there are theorists who claim the relevant kind of control is the ability to do otherwise. On such accounts, our agential features endow moral responsibility because (when suitably deployed) they empower us with a *freedom of choice* (or 'leeway') amongst alternative possibilities. Reasons-responsiveness theories have often been developed with this sort of control

[21] Shoemaker (2022) has recently set out a diversified conception of control, taking in evaluative and empathetic (as well as voluntarist) control. He thereby takes control to refer to more than just intentional action – it also includes evaluative judgments and interpersonal (dis)regarding. Shoemaker's intention is similar in spirit to mine: to reconceive control so that it can account for a fuller range of things which one can be morally responsible for.

[22] Although it is possible for a theory to specify a combination of agential features as essential for explaining moral responsibility. I suggest Susan Wolf's (1987, 1990) *sane deep self* view is an example of this, as I discuss in Section 1.4.4.

in mind. Reasons-responsiveness, as discussed, is a capacity which enables us to distinguish amongst available options and to select the most rational. The kind of conditional freedom described by Hobbes is also a clear expression of an ability to do otherwise. Much of the traditional debate around free will and moral responsibility centred on the plausibility of such leeway control, particularly in the face of determinism.[23] However more recently a different kind of control has taken centre stage. This is control in the sense of being the proper *source* of what one does. Real Self theories typically aim to articulate, and defend, control of this second sort. Indeed, Frankfurt's own theory of this kind was developed precisely in response to the perceived inadequacy of accounting for moral responsibility in terms of alternative possibilities, something for which he famously and separately argued (in Frankfurt 1969). I think libertarian views are also best understood as articulating versions of sourcehood control. John Fischer (2006: 51) refers to sourcehood views as "actual sequence" views.[24] The central idea here is that our agential features confer moral responsibility because they ensure our actions and attitudes emanate from the appropriate place. So, what matters for moral responsibility is how our actual agential features are exercised in relation to our actions and attitudes, such that we can be said, in some sense, to endorse or properly author these things. We need not learn about the availability of alternative possibilities with respect to our actions or attitudes – only their source or origin in the relevant agential features – in order to establish moral responsibility for them.

As I discuss in Chapter 4, it is control of the second variety – sourcehood – which has become the preeminent concern when it comes to skepticism about moral responsibility.[25] Likewise I will focus on skepticism of this variety in that chapter's discussion of the manipulation argument.

I accept that some may remain unconvinced that control is a concept which we can understand quite so broadly. In particular, some authors have understood sourcehood accounts of moral responsibility – such as quality of will theories – to represent an

---

[23] Peter van Inwagen's Consequence Argument was a particularly influential skeptical intervention into this debate (1983: 182–88), along with Frankfurt (1969) as mentioned.

[24] As far as I can tell, the *actual sequence* terminology originates with Susan Hurley (2003).

[25] Though of course there remains substantial debate over the viability of control in the sense of alternative possibilities and its necessity for moral responsibility – not least in the extensive literature spawned by Frankfurt's original 1969 argument against said necessity.

alternative to control-based accounts (rather than being themselves just another kind of control-based account). For example, consider the terms in which Chandra Sripada (2017) discusses Frankfurt's Real Self account. Frankfurt (1971: 12-20) famously distinguishes between a Willing and an Unwilling Addict, using this distinction to illustrate his conception of moral responsibility. The Unwilling Addict is he who lacks any second-order volition that his desire to take drugs be effective in motivating him to act (in fact, his second-order desires are to *refrain* from acting on his addictive desires). Thus, for Frankfurt, the Unwilling Addict is not morally responsible with respect to his drug-taking, since he lacks the requisite volitional structure. Whereas the Willing Addict *does* desire for his addictive desire to take drugs to be effective and cause him to take drugs. Thus, on the terms of Frankfurt's theory, the Willing Addict *is* morally responsible with respect to his drug-taking. He exhibits the kind of sourcehood – a Real Self constituted of appropriately congruent higher-order volitions – which Frankfurt identifies as essential for moral responsibility.

For Sripada, the kind of sourcehood which the Willing Addict exhibits, and which (according to Frankfurt) explains his moral responsibility, should not be understood as a form of control. Rather, it is a form of "self-expression", which is to be contradistinguished from "control-based" explanations of moral responsibility (2017: 781).[26] Others however take a different view to Sripada. They, like me, view sourcehood – such as quality of will in the Frankfurtian sense – to be a distinctive form of control of its own. Repko Waller puts this explicitly: "Frankfurt (1971) himself turns to sourcehood as the *kind of control* one must have over one's action to be held morally responsible" (2023: 305, italics added). On this reading, both the Unwilling and Willing Addict entirely lack leeway control – neither can do other than to take drugs. But the Willing Addict retains control in the form of sourcehood – he exercises agency over his drug-taking in a way the Unwilling Addict does not. Being the appropriate source of one's actions in this way is clearly a very different kind of thing to possessing the ability to do otherwise with respect to them. But I maintain it is

[26] Frankfurt himself may well have understood his account to be in opposition to control-based theories. But we ought to remember that Frankfurt was interested in debunking control *of a specific sort* as necessary for moral responsibility: the ability to do otherwise (or leeway control). This, as discussed, was how the free will debate had traditionally conceived of the kind of control required for moral responsibility. It is much less clear however that Frankfurt would understand his view to stand in opposition to control of *all* kinds, including the notion of control as sourcehood which has gained prominence in more recent literature.

reasonable – or at least it is not obviously unreasonable – to see both as different senses in which the control condition for moral responsibility might be satisfied.

1.3.4. The Humdrumness of Moral Responsibility

The various agential features highlighted by theories of moral responsibility have something else in common, in addition to being vehicles of control (liberally interpreted). They are each composed, at least partly, out of mental states and events such as the agent's intentions, desires, beliefs, judgments, decisions and values. To be reasons-responsive, or have a quality of will, or exhibit any of the other agential features discussed so far, is to form or exhibit one or more mental states or events of this sort. Different mental states or events may be implicated in different ways in different agential features. (For instance, Frankfurt's notion of higher-order endorsement relies only on desires, whereas Watson's version of this agential feature incorporates values as well. Voluntariness requires only the formation of intentions. Reasons-responsiveness, at minimum, involves forming judgments and beliefs in response to reasons). But we invariably find that mental states or events of one kind or another figure in theories of moral responsibility. The explanation, I suggest, is that the agential feature(s) upon which these theories are respectively premised are ones which supervene on mental states or events. Theories of moral responsibility all dress the skeleton concept of agential features in this same way, at least.

Accounts of agency have typically explained agency in terms of ascribing requisite intentional mental states with representational contents (Schlosser 2019b). This might imply that representational mental states are similarly prerequisite for my notion of agential features. Indeed, the features which seem to be most clearly *non-agential*, in this sense, are those which are non-mental. Consider eye colour or blood type, for instance. These are features which intuitively do not endow agents with control of *any* kind over their actions or attitudes. If it was proposed that an agent was morally responsible in virtue of possessing a particular eye colour or blood type, this would strike us as unusual and unpromising as a theory of moral responsibility. This is because purely physical, non-mental traits such as these are intuitively not the kind of thing which can be the locus of the control necessary for moral responsibility. Whereas theories of moral responsibility have unanimously held that mental states and events are different, and apparently unique, in their capacity to confer the relevant control.[27]

Perhaps it is nonetheless possible for agential features to be constituted out of entirely non-mental properties. Forms of agency have occasionally been proposed which do not require the ascription of representational mental states. For instance, Dreyfus discusses the idea of "skilful coping", a form of agency which does not involve conscious deliberation, reasoning or planning. Instead, we experience a "steady flow of skilful activity in response to one's sense of the situation", bypassing any representational mental states (2002: 378). To the extent that such proposals have merit, they may provide the resources to account for control-conferring agential features which do *not* rely upon mental properties. I leave this possibility open. For now, I simply observe that these have not been the kind of agential feature which have figured in theories of moral responsibility. Rather, it is exclusively mental states and events which have been deemed by theorists to be able to constitute the kind of agential feature we need for moral responsibility.

This points towards a final (and less often remarked upon) characteristic of those agential features which we find in theories of moral responsibility, which is worth registering. The fact that these agential features are built out of mental states and events means they are, in a sense, *humdrum*. The mental phenomena in question – intentions, desires, beliefs, judgments, values – are certainly all very familiar. They are garden-variety mental states which most mature adults, at least, regularly experience. By the same token, the agential features which supervene on these mental phenomena share this humdrumness. They describe kinds of control which few would question as not merely real, but commonplace.[28] So, whilst agential features purportedly have the power to elevate us into the realm of the morally responsible, they are also humdrum in this sense.

[27] Of course, many non-mental capacities (such as visual acuity or hand-eye coordination) *do* afford us greater control over what we do. But again, such non-mental capacities are not the kinds of agential feature which theorists have thought to be relevant for moral responsibility.

[28] This is true even for those who – as discussed in Section 1.3.3 – would resist interpreting all the agential features in question as versions of control, as I have proposed we can. What remains widely accepted, at least among philosophers, is that the agential features in question – quality of will, reasons-responsiveness, and so on – are real. (Libertarian agential features are the exception here, as I discuss in the main text).

This humdrumness accounts for how the agential features in question (excluding the libertarian kind) are uncontroversially able to figure in compatibilist accounts of moral responsibility. They, and the moral responsibility which purportedly depends upon them, remain feasible even if we posit a fully deterministic world, where all of our actions and attitudes are deterministically necessitated to occur. Similarly, it is their humdrumness which explains why even skeptics about moral responsibility tend not to deny that agential features exist. Rather than doubting whether we can control our actions in these ways, skeptics question the *significance* of the control which these agential features impart and its adequacy for moral responsibility.[29] The exception here are those agential features prized by libertarian accounts. Far from being humdrum, these theories – given their incorporation of indeterminacy – are somewhat exotic. This has led skeptics about libertarianism, in particular agent-causal libertarianism with its commitment to the existence of indeterminate agent substances, to challenge the plausibility that these agential features even exist (see Pereboom 2014: 65-60).

1.3.5. Instrumentalist Theories: The Exception to the Rule?

There is a prominent family of theories in the literature which I have not yet discussed: *instrumentalist* theories of moral responsibility. I deem these – at least one version of them – to be an exception to my claims about the essential connection between agential features and moral responsibility. According to instrumentalist theories – at least the substantial majority of them – what determines your status as morally responsible is whether *treating* you as such will promote desirable consequences. These views are instrumental since they view the means by which we hold ourselves and others morally responsible – our practices of blame and punishment, praise and reward – as instruments whose use can be justified by the good outcomes they secure (Milam 2021).[30] Instrumentalist accounts are therefore intimately tied to consequentialist moral theory in their focus on valuable outcomes.

[29] I discuss this further in Chapter 4.

[30] I recognise that *blame* and *praise* remain somewhat vague notions (at least I have always found them to be so). Without further explication, it is ambiguous what exactly these responses are. My treatment of blame in Chapter 2 seeks to address this, making explicit how I am choosing to understand these crucial concepts.

There are two broad categories of instrumentalist theory – direct and indirect. According to direct instrumentalist theories, we should ascribe moral responsibility wherever doing so promises to achieve desirable results. So, these theories aim *directly* at the desired consequences – we treat each other as morally responsible *in order to produce the good effects*, where the aim is to induce behavioural change. Moritz Schlick and J.C.C. Smart are two early, well-known advocates of the direct instrumentalist view.[31] Schlick offered a descriptive account: he thought that ascriptions of moral responsibility are in fact already made according to whether the ascription (and subsequent reward or punishment) will influence the agent as desired. He writes "[t]he question of who is responsible… is a matter only of knowing who is to be punished or rewarded, in order that punishment and reward function as such – be able to achieve their goal" (1939: 61). Smart (1961) instead offers a prescriptive thesis – he thought this *should* be the basis on which we ascribe moral responsibility. He argued that we ought to blame in order motivate the recipient to behave differently in the future. Crucially, for both Schlick and Smart, one's susceptibility to being incentivised in this way fully accounts for one's status *as* morally responsible.

What is the relevant agential feature which explains one's moral responsibility, according to these views? I do not see one. On these accounts, only one factor determines and explains whether or not one is morally responsible: whether or not treating them as such will promote our desired ends. Direct instrumentalist views therefore place no importance on whether the recipient of blame and praise possesses one agential feature over any another. The only feature demanded of the agent is their *influenceability* in being held morally responsible, via blame and praise. This influenceability *cannot* be considered an agential feature in the sense I have described. If one is influenceable, this does not entail they wield any kind of control over their actions or attitudes for which they are being held responsible (nor is it claimed to by the theorists who advocate these views, as I interpret them). My conception of control is inclusive, but not so inclusive that it interprets influenceability as a control-conferring feature. The absence of an essential agential feature, I suggest, means that direct instrumentalist theories cannot be deemed to be theories of moral responsibility

[31] Arneson (2003) is a relatively rare contemporary advocate of such an approach. Pereboom (2021a) is a leading skeptic about moral responsibility (in the basic desert sense), but he embraces a direct instrumentalist justification of our moral responsibility practices.

in the sense I am conceiving of the concept. I do not wish to aggressively police the boundary around theories of moral responsibility – I am not sure it is desirable or possible to set such limits. But regardless of what labels we choose, we should recognise that direct instrumentalist theories are describing a categorically different phenomenon than those others I have considered.

I think we can understand *indirect* instrumentalist theories of moral responsibility – the more popular variant in recent times – as, in effect, seeking to address this point. These views have spearheaded a recent resurgence in instrumentalist theories, but with an important twist. Unlike their direct counterparts, these theories *do* pinpoint and apparently privilege a specific, stable agential feature which must be present if moral responsibility is to obtain. The most prominent of these accounts cite reasons-responsiveness (or something akin to it) as the relevant capacity.[32] These accounts commonly claim that we should treat people as morally responsible – submit them to blame and praise – only if they are responsive to moral reasons. The rationale for this is still steadfastly instrumentalist: that such a practice will deliver the best overall consequences. Broadly, these accounts share the idea that adopting a practice of blaming those who are morally reasons-responsive will serve to cultivate or "scaffold" that very capacity of reasons-responsiveness, which is taken to be a morally desirable end (McGeer 2015: 2638; Jefferson 2019: 559). So, unlike with direct theories, we can identify the morally responsible, and explain their status as such, with reference to an essential agential feature (namely that of reasons-responsiveness). The overall practice of blaming and praising is justified by its positive scaffolding effects. But the norms by which we hold each other morally responsible refer only to this agential feature: "the *evaluation of the agent's action* gets settled by looking backwards" (Jefferson 2019: 562, original italics). This brings such theories closer, at least in terms of appearance, to theories of moral responsibility as I am characterising them.

However, I do not consider these indirect instrumentalist accounts to be theorising about the same phenomenon as the theories of moral responsibility considered in Section 1.3.1. This is because indirect instrumentalist theories fail to accord agential features the same kind of normative significance as those other theories. In indirect instrumentalist theories, the role of agential features in explaining the status of moral

[32] For instance: Vargas 2013; McGeer 2015; McGeer & Pettit 2015; Jefferson 2019.

responsibility is fundamentally different, and no longer irreducible. The explanation of one's moral responsibility *is* grounded, at least partly, in agential features. But the explanatory work is ultimately being done by something else: the fact that our moral responsibility practices promote valuable consequences. The significance of any agential control which figures in these accounts – whether in the form of the ability to do otherwise or sourcehood – is derivative. It is still, as it is in direct instrumentalist theories, the agent's *influenceability* which bears the real explanatory weight. (It is just that the influenceability is now *mediated* by an agential feature – reasons-responsiveness – which it is argued must be present for the influenceability to obtain). The following consideration helps us to see this. Were we to discover that a different practice would better secure the desired overall consequences – a practice which ascribed moral responsibility, blame and praise in response to the presence of a *different* agential feature than the target's reasons-responsiveness – then the consequentialist moral logic of such theories will recommend we adopt that practice instead. So, on these indirect consequentialist theories too, the agential feature of reasons-responsiveness is not *really* being privileged for its unique power to explain one's moral responsibility. Rather, it is being *exploited* for its utility within a practice which – contingently, in the current context at least – is efficacious in influencing us in the relevant desirable ways.[33]

A final point is worth noting here. Instrumentalist theories of moral responsibility, especially the earlier (Schlick/Smart) direct versions, have focused predominantly on

[33] Vargas' (2013) theory is something of an outlier. As a crude summary, it offers an instrumentalist practice-level justification of blaming and praising (in terms of fostering in the blamed agent "a kind of agency sensitive to and governed by moral considerations'' (2013:173)). But this practice is best served if individual instances of blame and praise are not justified with the instrumentalist justification in mind – they instead respond purely to whether the agent is reasons-responsive. This much Vargas' account broadly shares with those indirect instrumentalist views already considered. But Vargas' view is also distinctive because it sets *distinct* criteria for the status of moral responsibility. An agent is *morally responsible* if they are reasons-responsive, *regardless* of whether our blaming and praising practice serves to cultivate that agent's moral agency (2013: 199-233). Vargas' position has come under criticism from his fellow indirect instrumentalists, who question whether it is tenable to maintain this delicate separation between the conditions of morally responsible agency, and the conditions which justify ascribing blame and praise (McGeer 2015: 2642-2649; Jefferson 2019: 568). As we have seen, those authors instead insist that one's influenceability *is* precisely what renders one a morally responsible agent, and not merely what justifies blame and praise. I am sympathetic to their critique of Vargas' account, but I cannot explore this further here and will not take any official stand. I only note that, to the extent that Vargas' account succeeds, it *would* satisfy the agential features requirement for a theory of moral responsibility as I understand it. It affords agential features an irreducible role in explaining the status of moral responsibility, not one which is contingent upon an agent's influenceability.

the influenceability of the *blamed agent,* and how this can be manipulated to achieve desirable results. I have described influenceability in similar terms so far. But the consequentialist moral logic underpinning these theories is one which presumably measures the rightness of a response with reference to *all* its consequences. If so, then instrumentalist accounts ought not to be interested merely in how the recipient of blame can be influenced. They ought to be interested how our blame and praise can influence *all future outcomes*. The influence of our blame and praise on the blamed agent – improving their behaviour, cultivating their moral agency – concerns but one set of outcomes amongst many others. Jefferson, in particular, has recognised these potential wider instrumental benefits. As she says, "I endorse the view that the effect of our moral responsibility practices on *all* participants contributes to justifying them" (2019: 559, original italics). Here she is referring to how our moral responsibility practices can also serve to scaffold the moral agency of third parties and the moral community as a whole.[34] These considerations, I think, only serve to reiterate the diminished role which the blamed party's agential features play in explaining their moral responsibility, in such instrumentalist accounts (both direct and indirect). The blamed agent may well be *particularly* positively influenceable via blame, compared with third parties, in virtue of their possessing agential features (typically reasons-responsiveness). But the interest in their influenceability is but *part* of a broader interest in influencing outcomes more generally for the better. This, again, is why I understand instrumentalist theories to be in the business of explaining *moral responsibility* in a categorically different sense to those non-instrumentalist theories of moral responsibility discussed earlier.

1.4. Blameworthiness (and Praiseworthiness)

I have set out the first component of my overarching conception of moral responsibility. This establishes it as that kind of responsibility which depends upon our possession of agential features, which are any feature which endows a person with some form of control over their attitudes and actions. I have observed it is exclusively, and consistently, *mental* states and events that theories of moral responsibility have deemed able to provide the requisite control. But, if possessing agential features of

[34] Jefferson (2024) has also argued that blaming the dead can be warranted on similar, indirectly instrumentalist grounds, citing possible benefits to the contemporary moral community.

this kind is irreducibly essential in explaining moral responsibility, the next question is why should the control which they confer be so essential? Why are agential features the common currency of theories of moral responsibility? This brings me to the second component of my overarching conception.

1.4.1. Moral Responsibility and Response-Worthiness

My initial answer to the questions just posed is that agential features, in virtue of the control which they endow, are thought to have a very particular kind of normative significance. They are essential for the generation of a unique kind of normative reason or consideration in favour of a particular kind of response: blame and praise.[35] Wherever this reason exists, it permits describing the agent in question as either *blameworthy* or *praiseworthy*, respectively. So, without possessing agential features, an agent cannot attain the status of blameworthiness and praiseworthiness. As I sought to emphasise in Section 1.2.1, causal responsibility very often does not entail any response-worthiness. And whilst assigned responsibility very often does, it does not necessarily do so. Moral responsibility is different – it is a status which has a distinctive form of response-worthiness *built into it*.[36] Certain kinds of reason, to respond in a certain kind of way – standardly referred to as blame and praise responses – come into effect.

The task of this section is to give some account of what this distinctive kind of response-worthiness involves. I will mostly restrict my discussion to blameworthiness and blame, rather than praiseworthiness and praise. This is for reasons of efficiency rather than any assumed philosophical differences between the two. So, unless

[35] My use of *normative reason*, and *consideration in favour*, is intended to be maximally inclusive with regards to the normative quality of the reason or consideration. So, for example, considerations of rightness, goodness, virtue, desert, merit, fit, aptness – and so on – would all fall under its umbrella. Some of these normative considerations may therefore not necessarily be directly available in our deliberative weighing of reasons. For example, the fittingness of an emotional response (fear, say). Assuming we understand emotional responses to be involuntary, such fittingness is not a consideration we can refer to in deliberating about how to respond. But its fittingness nonetheless represents how the emotion is *normatively favoured* by certain facts (such as the presence of danger). It is therefore in scope and qualifies as a normative reason as I am using the term.

[36] At least, as I will discuss, it has this inherent response-worthiness in the case of moral responsibility for anything which is not morally neutral. For morally neutral things, response-worthiness does not apply (though I think we might plausibly insist it still exists, just as latent or not activated, wherever the relevant agential features are instantiated). But in *all* cases of assigned responsibility, only the assignation of duties is essential and inherent; any response-worthiness is not inherent – it is something additional to the initial specification of the responsibility that is assigned.

otherwise stated, my comments on blame and blameworthiness should be taken to apply, by extension and symmetrically, to the praising side of the coin.

Clearly, no theory of moral responsibility posits that just *any* control-conferring, agential feature will do the job and possess the normative significance in question. Were this the case, the skeleton concept of agential features would already be sufficient to account for the agential feature requirement. Rather, as we have seen, theories of moral responsibility each tend to identify a particular, privileged agential feature which they take to bestow the *relevant* kind of control. Thus, any given theory of moral responsibility need not (and typically does not) deny that we are capable of possessing those agential features, and the associated kind of control, which are highlighted by its rival theories. (For example, a quality of will theorist can – and almost certainly will – accept entirely that we are capable of voluntariness and reasons-responsiveness). What they deny is that those rival theories have identified the *relevant* agential feature for explaining moral responsibility.[37]

But *how* is it that agential features contribute to explaining or accounting for moral responsibility? In the first instance, I think, because they confer morally responsible *agency*. As I understand it, theories of moral responsibility typically maintain that mere possession of the agential feature(s) which they distinctively highlight is both necessary and sufficient to qualify one as a morally responsible *agent*. For instance, if we favour a reasons-responsiveness theory, then merely possessing a functioning capacity for reasons-responsiveness already renders us morally responsible.[38]

However, as well as being morally responsible agents, in the abstract, all theories of moral responsibility agree that we can be morally responsible *for* specific things – doings or attitudes, say. These are the occasions on which we both can become blameworthy and praiseworthy, and where in responding to us as such, we are appropriately *held* morally responsible. So, to be a morally responsible agent (in virtue

[37] Or at least, they deny that those theories have identified the relevant agential feature *for generating that variety of moral responsibility which they are seeking to articulate*. As I discuss in Chapter 2, it is possible to be pluralistic about moral responsibility. More than one theory might be true. Different agential features might provide considerations in favour of different kinds of blame and praise, in which case we can be morally responsible, and blameworthy, in more than one sense. (Though as noted in the previous section, theories of moral responsibility will usually explain the moral responsibility in question in terms of a singular agential feature, rather than multiple).

[38] McKenna (2012: 10-14) and Shoemaker (2015: 5-6) give especially clear articulations of the notion of morally responsible agency.

of our agential features) is also to render us a *candidate* for something else: blameworthiness and praiseworthiness.[39]

1.4.2. Conditions of Blameworthiness I: Non-Neutral Moral Performance

This does not mean that for every action which a morally responsible agent performs, or for every attitude they hold, they are necessarily either praiseworthy or blameworthy with respect to them. One reason they might lack blameworthiness or praiseworthiness is if the action or choice or attitude in question is morally neutral. For example, if Parminder chooses to walk home via the park rather than the beach, this choice is unlikely to be a morally significant one. Most moral theories will remain silent on the moral status of such a choice, deeming it neither to surpass nor fall short of any moral expectations. In this sense Parminder's choice would be a morally neutral one: thus, not one with respect to which he can be blameworthy or praiseworthy. Supposing this is true, it can still be the case that Parminder was morally responsible for the choice, provided he possessed the relevant agential features as identified by our preferred theory of moral responsibility. For instance, his choice might still be a product of his exercising a robust capacity of reasons-responsiveness, or one which he fully endorses given his desires and values.[40]

In practice, though, it seems we only tend to say of agents that *they are morally responsible for X* in the case of *non*-neutral moral performance. That is, for blameworthy or – though less common – praiseworthy performances. The concept of moral responsibility is especially useful, it seems, in such cases. Imagine a (somewhat dull) world in which everybody performed only morally neutral actions, such that no praiseworthiness or blameworthiness ever arose. They may yet satisfy the conditions for being morally responsible for what they do, in virtue of their possessing the requisite agential features, but it is difficult to see what role the concept of moral responsibility could usefully play. I suggest that moral responsibility gains much of its utility as a concept because we happen to be creatures who perform either successfully

[39] The titles of the upcoming Sections 1.4.2 and 1.4.4 – 1.4.6 refer only to blameworthiness. Not to belabour this point, but I do consider my arguments to have extension to praiseworthiness too, as my discussion indicates. But my focus, as throughout the thesis, remains foremostly on blame and only secondarily on praise.

[40] Zimmerman (1988: 61-62) offers insightful comments on moral responsibility for morally neutral actions.

or unsuccessfully relative to moral standards or norms. So, we can say it is valuable in virtue of its response-worthiness – its distinctive and inherent power to license responses of praise and blame.

So, for an agent to attain either the status of blameworthiness or praiseworthiness – for these distinctive reasons to blame or to praise to come into existence – the mere possession of agential features is never enough. Agential features must also be accompanied by a *non-neutral moral performance*.[41][42] In the case of blameworthiness this is moral performance which is in some way substandard. Wrongful behaviour is the paradigmatic example – it falls below a generalisable standard of moral justification. Unprovoked violence towards fellow pedestrians, say, is generally viewed as a clear failure to meet our moral obligations. But we can morally underperform in other ways – suberogatory actions for instance (perhaps staring angrily at pedestrians would fall into this category). Or in our failures to meet the more stringent expectations to which only some – close friends or partners, say – hold us, but to which others do not. These too are moral underperformances which theories of moral responsibility have often deemed we can be blameworthy for, and for which we do in fact observe blame being doled out.[43] So, the scope of blameworthiness might plausibly extend beyond the strictly deontic category of wrongful acts (McKenna 2012: 173-74). It encompasses other kinds of disappointment with regards to moral

---

[41] This is why earlier I suggested that we might consider that the response-worthiness remains latent in such cases. By which I mean it is still there, inherent within *all* instances of morally responsible behaviour (including Parminder's choice of the park). It just remains 'untriggered' due to the moral neutrality of the choice. (Were the moral status of the choice to be different – perhaps we come to change our moral principles – then the exact same act, involving the exact same agential features, might become non-neutral. The response-worthiness inherent to those agential features would then be triggered, in the form of either blameworthiness or praiseworthiness).

[42] The requirement of *morally* non-neutral performance distinguishes blameworthiness, and moral responsibility more generally, in the sense I am interested from their putative equivalents in other domains. I said earlier that a tri-partite distinction between causal, assigned and moral responsibility is traditional. However, increasingly some authors have proposed there are such things as epistemic, or aesthetic, or perhaps even athletic, responsibility (Shoemaker 2015: 5). If so, then I propose that blameworthiness or praiseworthiness in these domains will analogously be achieved in virtue of performance which is non-neutral with regards to epistemic, aesthetic and athletic standards, rather than moral ones. (We may have a reason to lavish Matisse with *aesthetic* praise for his artworks, despite knowing nothing of his, or his artworks', moral qualities). Whether or not agential features are also essential for responsibility and blameworthiness in these other, non-moral domains however is something I leave open. For helpful discussion and comparison between these other kinds of blameworthiness see Kubala (2024) on aesthetic responsibility, and Matheson & Milam (2021) on non-moral blame more generally.

[43] Wallace (1994: 37-38) is especially instructive on these different kinds of moral expectations which might be relevant for blameworthiness.

expectations – behaviour which is reasonably thought to be morally subpar in the eyes of the blamer, according to standards to which they hold the target of blame.[44]

With praiseworthiness the opposite is true – the possession of agential features must be supplemented with a *surpassing* of moral expectations. For instance, refraining from violence towards fellow pedestrians meets a widely accepted moral obligation, but does not typically garner praise. Rushing to the defence of a person being subjected to such violence, however, might well go beyond what could be reasonably expected – and we might well expect this supererogatory act to elicit praise. This symmetry in terms of the moral performance condition for blameworthiness and praiseworthiness is intuitive, and it seems to be standardly accepted. As with blameworthiness, the relevant moral expectations might not necessarily be purely a function of the deontic status of the moral performance. Different people might be held to different expectations; if doing the right thing is circumstantially difficult, praise *might* be owed merely for doing what one is morally obligated to do. (For instance, if you find a briefcase full of cash, we might consider you praiseworthy for handing the money in, even if that is precisely what morality dictates you ought to do).

What is the range of things to which these moral expectations pertain – what is the scope of things we can be morally responsible *for*? Actions, and omissions, and also the consequences of both – this much is shared with the typical scope of assigned responsibility (see Footnote 2). But the objects of moral responsibility are often thought to range more widely still: that we can be morally responsible for our *attitudes* and even our characters (Smith 2005). These too can be in suitable relationship with our agential features such that we are morally responsible for them (and thus either blameworthy or praiseworthy for them, where they are morally non-neutral).[45]

---

[44] Of course, these expectations must in some sense be legitimate. It is well outside the scope of this thesis to address which of our moral expectations are legitimate, as this will largely rest upon what one takes the correct moral theory to be. I simply assume that legitimate or reasonable moral expectations, in relation to various moral standards (such as standards of wrongfulness and suberogation; as well as relationship-specific standards) can and do exist.

[45] My talk of morally subpar *behaviour* or *performance* is therefore generally meant to include each of these objects, and not merely the doings or actions which they most naturally suggest.

1.4.3. Excuses and Exemptions

I have just said there must be a suitable explanatory connection between the relevant agential feature and that which we are putatively morally responsible for, if we are in fact to be morally responsible for that thing. What makes for a *suitable* explanatory connection? This remains obscure, for now. But we can begin by noticing those situations where it is clearly *lacking*.

It is at least clear that the mere *coincidence* or co-occurrence of our agential features and our actions (or attitudes) is insufficient to render us morally responsible for those things. This is demonstrated by cases of excuse. For instance, suppose Erica pushes Elijah, but only because her family was threatened with harm if she did not. Most theories of moral responsibility will excuse her from (or at least substantially mitigate her) blameworthiness. For instance, she is inhibited from responding to moral reasons as she normally would. Her real self does not endorse her action. It does not betray her ill will towards Elijah. In such an excusing case, Erica maintains her status as a morally responsible agent (in virtue of her agential features) but fails to be morally responsible or blameworthy *for* the thing in question. The coercive context within which she makes her decision to push Elijah robs her agential features of whatever power they would otherwise have to endow her with blameworthiness-conferring control over that decision. We are *exempt* from blameworthiness, on the other hand, where we simply do not possess the relevant agential feature to begin with. Suppose the unfortunate Elijah is now pushed by his toddler sister Edna. Toddlers are normally understood to lack the agential features required for moral responsibility – reasons-responsiveness, higher-order desires, quality of will and the like.[46] Edna is therefore exempted. She is not a morally responsible agent and thus (unlike Erica) is not even a *candidate* for blameworthiness.

I suggest that the notion of agential features, then, provides an illuminating prism for thinking about cases where agents are excused or exempted from blameworthiness. These are cases where an agent does not satisfy the conditions for blameworthiness, even whilst they do (at least arguably) morally underperform. The reason is that their

[46] Strawson (1963: 204), as one example, includes very young children among those who are not eligible for the blaming emotions, since they cannot display the requisite poor quality of regard.

agential features are either absent or are inhibited from playing a suitable explanatory role in their moral underperformance.[47]

1.4.4. Conditions of Blameworthiness II: Deployed Agential Features

My discussion of excuses reveals how, for one to be morally responsible *for* something – and thus either blameworthy or praiseworthy – the relevant agential feature must in some sense be *deployed* in the occurrence of that thing. What does this deployment entail? When is it that agential features *are* empowered to (purportedly) endow the control which explains blameworthiness and praiseworthiness?

A full answer will depend on the agential feature in question, and should be left to those theories of moral responsibility in which it figures to articulate. But I think we can identify at least two broad senses in which agential features can be deployed that have been thought to make for a suitable explanatory connection. First, where the object of moral responsibility is in a *causal* relationship with the agential feature, such that it results from it. For instance, reasons-responsiveness plays its suitable explanatory role where the decision we take is caused by a deliberative process which is reasons-responsive.[48][49] Second, a relationship of *endorsement*, where the object of moral responsibility is endorsed by or expresses the relevant agential feature. The various Real Self theories describe this form of relationship – our agential features comprise our underlying character, which endorses those attitudes or actions for which

[47] This distinction between excuses and exemptions, now common in the literature was (I think) first clearly stated by Watson (1987). Brink and Nelkin (2013: 291) persuasively argue that the distinction between excuses and exemptions is not a bright line. Rather, they exist along the same spectrum, with exemptions being effectively conditions of global or permanent excusal.

[48] This does not mean that the reasons-responsiveness capacity must be exercised *well* if there is to be moral responsibility. It might be exercised poorly – or indeed fail to be exercised at all where it *ought* to have been. (As Shoemaker puts it with regards to the latter: "quality of judgment does not have to be restricted to actual judgments; rather, it could merely be about missing judgments, where one's decisions were at least governable by judgment" (2015: 67)). These are the very two situations in which there is blameworthiness according to reasons-responsiveness accounts.

[49] Following on from the previous footnote: whilst I have argued that agential features are composed out of, or otherwise implicate, mental states and events, they need not take the form of instantiated or active mental states at the time we perform an action or adopt an attitude. On some accounts, what matters instead is that the agent possesses a *capacity* to have particular mental states. Reasons-responsiveness theories are again notable in this sense. On reasons-responsiveness accounts, what matters is that we possess a general capacity (to form beliefs based on existing reasons). As discussed in Footnote 48, exercising this capacity unsuccessfully might sometimes involve failing to be reasons-responsive at all when we might have been. This failure will therefore manifest in an *absence* of the relevant mental states (beliefs based on good reasons).

we are morally responsible.[50] I would argue a third, distinct possible form of relationship is *constitutive*. Here, the agential features need not be accompanied by a specific decision or action or attitude of which they are the cause, or for which they offer endorsement. Rather, the agential features themselves are the object of moral responsibility. This would apply perhaps most clearly to those theories which posit we can be morally responsible for our characters – that we can be blameworthy or praiseworthy for the vicious or virtuous quality of our character (or our real self) itself. Robert Adams (1985) is a notable proponent of such a view.[51]

So, we at least have an idea of what it means for agential features to be deployed in the relevant sense. This does not mean that the agential feature must necessarily be deployed at the time of the moral underperformance for which we are blameworthy. For instance, suppose Elton chooses to ingest an elixir which he knows will cause him to experience an involuntary, violent spasm of the legs a few days later, when he will attend a crowded music festival. Assume his choice was one in which his agential features were deployed (his real self endorsed his choice and he was fully receptive and responsive to moral reasons, say). Later, at the festival, Elton's legs are predictably (at least for him) sent into paroxysmal frenzy. He injures several of his fellow festivalgoers. No agential features are deployed in his involuntary spasm, a purely physical (non-mental) event. But because the spasm is causally *traceable* to his earlier choice – in which his agential features *were* deployed and for which he is (let us assume) blameworthy – Elton would be blameworthy for it.[52]

[50] In such *endorsing* cases our agential features will still typically be in a causal relationship with the object of our moral responsibility. The action for which we are morally responsible, say, is the causal product of those mental states and events which comprise our agential features, such as desires and beliefs. But here it is not the causal power of the agential features which does the explanatory work. It is their power to endorse our actions or attitudes – on behalf of our real self – which explains how these agential features endow moral responsibility.

[51] Again, instrumentalist theories do not seem to fit comfortably within this framework. It is difficult to see how the feature of the agent which they cite as irreducibly essential for their moral responsibility – their influenceability – is *deployed* in any such way. Certainly, our influenceability via blame and praise for our actions or attitudes does not entail that we cause, nor constitute, nor endorse, those actions or attitudes. Rather, it would seem to *exploit* them, after the fact. If I am susceptible to influence via blame and praise for an action I have performed, say, this does not seem to offer any *explanation* of that action's performance, much less one which explicates how we exert control with respect to it. Again, I think this should push us towards seeing instrumentalist theories as identifying categorically different species of 'blameworthiness' and 'praiseworthiness'.

[52] The idea that one's moral responsibility can be *traced* in this way is now fairly common. Fischer and Ravizza (1998: 49–51) and Vargas (2005) give useful sustained discussions of this idea.

A final wrinkle to note. I discussed earlier how there is a standardly assumed symmetry in terms of the moral performance required for blameworthiness and praiseworthiness. Blameworthiness tracks underperformance and praiseworthiness tracks overperformance in relation to moral expectations. It is most often assumed that there are similarly symmetrical conditions when it comes to the kind of control or freedom which is required for blameworthiness and praiseworthiness. Or, as I would put this, whether the same *agential features* must be deployed in each case. Some however have developed theories of moral responsibility which set asymmetrical conditions here. Perhaps most notable is Susan Wolf's (1990) account. Wolf is a Real Self theorist. She argues we are free and morally responsible with respect to our actions wherever they reflect the values and commitments which comprise our real self (1990: 85–9). Praiseworthiness requires only that we possess such a real self. But she adds an additional, more stringent condition for blameworthiness. It also requires a capacity to understand the difference between moral right and wrong – or (as she puts it) the "True and the Good" (1990: 117) – so that we might have reasonably avoided wrongdoing.[53] In addition to this asymmetry, this theory is also notable and unusual in another way. It is an example of a theory of moral responsibility which demands (at least for blameworthiness) a *combination* of deployed agential features: a real self, and a form of reasons-responsiveness.[54]

1.4.5. Epistemic Features and Blameworthiness

Are there any other facts – aside from the control bestowed by agential features, when they are suitably deployed in disappointing moral performance – which are essential for generating a reason of blameworthiness? One plausible contender is a fact about our *epistemic* status. An epistemic condition on moral responsibility is standardly accepted alongside the control condition (the latter I have cashed out in terms of agential features).[55] Roughly, the epistemic condition says that if we are to be morally responsible for our actions or our attitudes, we must know what our attitudes and

[53] Elsewhere Wolf (1987) refers to this capacity as *sanity*. Nelkin (2011) offers an asymmetrical account very similar in spirit to Wolf's, specifying that certain rational abilities are needed for blameworthiness but not for praiseworthiness.

[54] To be clear, other theories might implicate more than one agential feature in their account of moral responsibility. (For instance, being able to act voluntarily would seemingly be a pre-requisite for being responsive to reasons). But my impression is that most other theories, unlike Wolf's, privilege only one agential feature as relevant for *explaining* moral responsibility.

[55] Rudy-Hiller (2022) provides a useful discussion of the epistemic condition on moral responsibility.

actions *are*, and also have some grasp of their moral status. Ignorance of these things would therefore comprise another source of excuse.[56] If I fail to realise I am blocking your path – or if I do realise it, but (surprisingly) I fail to realise that this is a morally questionable thing to do – then either of these things would intuitively seem to mitigate my moral responsibility, and my blameworthiness, for my behaviour.

Plausibly, meeting this epistemic condition should also be understood as possessing a particular kind of agential feature, especially given the inclusive notion of control which I have advocated. Meeting it presumably involves having certain mental states: *beliefs* about our actions and attitudes which meet some standard of accuracy. And surely having awareness of what we are doing and of the content of our attitudes, as well as understanding their moral status, grants us some greater degree of control with respect to these things.[57]

I am therefore happy to accept that the epistemic condition on moral responsibility and blameworthiness (and praiseworthiness) can reasonably be understood as a requirement that the blamed agent possess a particular sort of agential feature. Meeting this requirement involves having a distinct kind of control over what we do. But I think it is helpful to keep this kind of agential feature distinct from the other control-conferring agential features I have described to this point. This distinction is reflected in the literature, which tends to address the control and epistemic conditions separately (with distinct lines of skepticism developed with regards to each). It is also the *control* condition which I will have particular interest in, as it is this (as opposed to the epistemic condition) which skepticism about basic desert moral responsibility has targeted, as I will discuss in Chapter 4. So, to reflect this distinction I will refer to *control-conferring* and *epistemic* agential features (whilst acknowledging that the terminology is imperfect, given that epistemic capacities would undoubtedly be aptly

[56] Again, provided that our ignorance is itself non-culpable. If our ignorance can be traced to an earlier decision for which we *are* morally responsible and blameworthy – perhaps we freely chose not to attend any of our undergraduate ethics lectures – then we might inherit blameworthiness for our future ignorance. It would no longer serve as an excuse.

[57] Indeed there is some debate about whether the epistemic condition collapses into the control condition. Mele (2010) argues for this, Montminy (2021) against. Whichever way this debate is resolved, these represent different *kinds* of control. I will therefore treat them distinctly.

described as *control-conferring*).[58] To be clear, though, epistemic features are included within the complete set of agential features.

1.4.6. Facts Which Influence Reasons to Blame but not Blameworthiness

In order to more sharply define the domain of facts which can influence blameworthiness, it is useful to briefly contrast the facts within this domain to those which sit outside it. In particular, with facts which *can* influence our reasons to blame, but which do not bear upon reasons of blameworthiness. One important such fact often discussed in the moral responsibility literature concerns having the requisite *standing* to blame. The most frequently cited conditions which can undermine one's standing to blame are complicity, meddling and hypocrisy. For instance, if a state is responsible for creating criminogenic conditions which predictably encourage criminal wrongdoing, then this might be thought to render the state complicit in that wrongdoing, thus undermining standing to blame and punish criminal wrongdoers (Tadros 2009). Alternatively, if two romantic partners are having a quarrel perhaps it is just not our business to blame one of the parties for their conduct, even if they could quite reasonably blame each other. To do so would be to meddle where we do not have the appropriate authority or standing (Seim 2019).

Both complicity and meddling concern facts about the relationship which the would-be blamer has to the recipient of blame. Hypocrisy seems slightly different. This instead is about the relationship of the would-be blamer to the relevant moral underperformance. We are hypocritical where we, say, have ourselves committed the same wrong. Some, such as Wallace (2010), argue that our own moral performances, if similarly disappointing, can remove our standing to blame others (if not ourselves).[59]

Certainly we find our actual responsibility practices reflect the apparent power of these standing-related considerations to affect blame's justification. We often see blame withheld because it would be hypocritical or because there is complicity (or at least

---

[58] Perhaps we can put it (inelegantly) like this: the *control-conferring* agential features which are my chief concern are any which endow control *other than the kind of control endowed in virtue of satisfying the epistemic condition on moral responsibility*.

[59] Standing to *praise* is much less discussed in the literature; both intuitively and in practice, worries about hypocrisy, complicity and meddling seem to be less threatening when it comes to praiseworthiness. See Telech (2024) for an interesting recent discussion about standing to praise, arguing that would-be praisers must themselves be committed to the relevant value, otherwise they lack standing to praise.

we think it should be suspended in such cases). But my sense is that these facts about standing should *not* be viewed as facts which affect the status of blameworthiness. They instead affect the status of something else: the all-things-considered justification of blame. I recognise this is a stipulation, and not something I claim to decisively establish.[60] Somewhere a line must be drawn between the conditions of blameworthiness, and the conditions supplementary to blameworthiness but relevant to the wider justification of blame.[61] I suggest the concept of blameworthiness is particularly *useful* if we use it to cordon off reasons to blame which are irreducibly grounded in the deployed agential features of the blamed agent. Considerations of standing (or lack thereof) do not impinge on these reasons – these are instead either relational facts or facts about the would-be blamer's own moral record.[62]

It is also important to note that, however we delimit them, blameworthiness and praiseworthiness are generally understood to provide only *pro tanto* reasons in favour of blame and praise. They are reasons which may be defeated in the all-things-considered analysis of whether or not blame or praise is morally justified. We might refrain from blaming the blameworthy, for example, where doing so threatens to so upset and destabilise them that they will endanger themselves or others. It would not (morally speaking) be worth blaming them, despite their blameworthiness (Smith 2007: 478-484). Conversely, there is at least conceptual space for the possibility that, all-things-considered, the non-blameworthy can justifiably be blamed. Calhoun for instance argues that blame of non-culpably ignorant (and thus non-blameworthy) wrongdoers in an oppressive masculine society might be justified where it is a means towards achieving social reform and, moreover, if failing to blame is likely to signal endorsement or sanctioning of the moral order (1989: 400-5).[63]

---

[60] Although it does seem to agree with the commonly held view in the moral responsibility literature. R.A. Duff's (2010) influential writings about standing to blame in *legal* contexts take a similar view.

[61] Angela Smith (2007) was one of the first to draw attention to the importance of making this distinction.

[62] Likewise, the reasons to blame defended by instrumentalist theories of moral responsibility are not reasons of blameworthiness and praiseworthiness in this sense. I have argued they are not reasons irreducibly grounded in an agent's deployed agential features (see also Footnote 51). On my interpretation, both direct and indirect instrumentalist theories are instead in the business of identifying all-things-considered reasons as to why we should engage in blame and praise.

[63] If we blame without *any* appraisal of blameworthiness, I am doubtful however whether we should conceive of this as genuine blame. My discussion of blame in Chapter 2 will bring this out.

1.4.7. Summing Up

This completes my conceptualisation of blameworthiness and praiseworthiness. These are statuses which each constitute a *pro tanto* normative reason of some description to respond in a certain way – with blame and praise. I have argued these normative reasons are generated out of a circumscribed set of facts, of which I have identified two essential members. First, non-neutral moral performance. In the case of blameworthiness, this means underperformance relative to moral expectations (in the case of praiseworthiness, overperformance). Second, agential features deployed in that non-neutral moral performance. Both of these facts are an ineliminable contributor to any reason of blameworthiness or praiseworthiness. I therefore propose we think of these facts as the two *essential constituents* of blameworthiness and praiseworthiness.[64]

When I use *blameworthiness*, then, the full expression of what I mean is *worthy of blame in virtue, at least partly, of agential features deployed in moral underperformance*. The 'at least partly' qualifier here is important. I have argued only that these are two essential constituents of any instance of blameworthiness and praiseworthiness – they are necessary, but they might not be sufficient. We might theorise, at least in principle, that these essential constituents are inadequate by themselves to generate blameworthiness and praiseworthiness. Perhaps they must be supplemented by further facts if they are to gain this reason-generating significance. These additional facts might be considered a supplementary *third* constituent of blameworthiness. Unlike the other two components, though, this constituent is *inessential*.[65] Indeed, my impression is that theories of moral responsibility – including each of those canvassed in this chapter – overwhelmingly do *not* call upon such additional facts to explain moral responsibility and blameworthiness. Real Self theories, for instance, typically refer only to one's possession of a real self in explaining moral responsibility. Quality of will theories more generally explain moral responsibility solely in terms of quality of will. The same with reasons-responsiveness

[64] I will occasionally hereafter use the phrase *deployed agential features*. It should be taken to mean agential features which are *deployed in non-neutral moral performance*, and thus describe either a case of blameworthiness or praiseworthiness.

[65] By which I mean it is not essential or inherent to the *concept* of blameworthiness. Particular theories of moral responsibility may yet deem such additional facts essential to their particular brand of blameworthiness (such as contractualist theories, as discussed in the text).

and libertarian theories. According to these theories, then, deployed agential features do *all* the normative work in generating reasons of blameworthiness.[66]

There is one kind of moral responsibility theory I have not yet mentioned yet, which is perhaps the notable exception here. These are *contractualist* theories. Scanlon (1998) has offered the most prominent version.[67] His account can be broadly understood as a quality of will account. It understands morally responsible agency and blameworthiness in terms of the quality of our judgment about reasons – or as Scanlon refers to it, the quality of our "judgment-sensitive attitudes" (1998: 275–276). But Scanlon is explicit that quality of judgment acquires this significance only given his broader contractualist moral theory. Summarised crudely, Scanlon's position is that a proposal for contracting into a practice which ascribes blame and praise according to the quality of judgment-sensitive attitudes could not be "reasonably rejected" (1998: 85). So here the deployed agential feature (quality of judgment) is *not* itself sufficient to generate blameworthiness. Additional facts *are* essential, namely the normative resources of contractualism. As McKenna puts it, "the normative warrant for blaming is, at least to some extent, taken out of the control of the free agent" (2024: 111).[68] This question of the sufficiency of deployed agential features for generating blameworthiness is central when it comes to analysing *basic desert* moral responsibility, as I do in Part Two of the thesis.

1.5. Concluding Remarks

The aim of this chapter has been to table the first two components of my overarching conception of moral responsibility. I have identified and refined these components by surveying some of the most prominent theories of moral responsibility in the literature. I have articulated the first of these components – agential features – extremely expansively. It describes those features of a person which confer upon them control of

[66] Or, if there *are* additional facts doing the normative heavy lifting in these theories, then (as I read them) these normative assumptions are not made explicit, and they ought to be.

[67] Lenman (2006) offers another well-known contractualist account of moral responsibility.

[68] We might think we can interpret instrumentalist theories in these same terms. Do these not similarly draw upon additional (consequentialist) normative resources to explain ascriptions of moral responsibility and blameworthiness? I am open to this interpretation (in which case these theories *do* articulate genuine reasons of blameworthiness and praiseworthiness, in contrast to what I said earlier). But my sense is that Scanlon's contractualist theory still cites the *deployment* of a particular, privileged agential feature as essential in explaining blameworthiness in a way in which instrumentalist theories (at least as I have argued) do not. This seems an important difference.

any description over their actions, attitudes or character. I have noted how theories of moral responsibility have consistently articulated those agential features they deem requisite for moral responsibility in terms of mental states or events.

The second component of my overarching conception is blameworthiness and praiseworthiness. These are statuses which generate normative reasons to blame or to praise, and account for the response-worthiness which is distinctively inherent to moral responsibility. Blameworthiness and praiseworthiness have two essential constituents: non-neutral moral performance, and the suitable deployment of agential features in that performance.

It is important to emphasise that my aims in this chapter have been mostly conceptual. I have not defended a normative claim that any of the deployed agential features I have considered here are in fact capable – whether independently or in tandem with additional normative facts – of generating blameworthiness. It is thus consistent with all that I have said (even if it seems very unlikely) that moral responsibility and blameworthiness do not exist.

There is a third and final component of my overarching conception of moral responsibility; one to which I have occasionally referred but said little about. This is the responses of blame and praise. I turn to this component next.

Chapter 2 - Conceptualising Blame and Introducing Basic Desert

2.1. Introduction

I now turn my attention to the third and final component of my overarching conception of moral responsibility. So far, I have offered an account of agential features, and argued that these are an essential member of a select group of facts which together are understood by theories of moral responsibility to generate the distinctive normative reasons of blameworthiness and praiseworthiness. Having established the minimal conditions which much be in place for these kinds of reasons to arise, I am now interested in the nature of the response which these reasons distinctively favour. So far, I have only established that the responses in question are *blame* and *praise*. But what do these responses involve? Until we have an answer to this question, it remains obscure what blameworthiness and praiseworthiness really amount to, and my conception of moral responsibility remains incomplete.

A perspicuous answer is not readily forthcoming from the moral responsibility literature, however. I have already discussed how the moral responsibility debate is marked by wide divergence in how its foundational concepts are understood and used. This is no less true when it comes to blame and praise. Theories of moral responsibility refer consistently to blame and praise; liability to these responses is clearly understood as integral to whatever it means to be morally responsible. Yet despite their centrality, the concepts of blame and praise are used by different theorists to denote very different things, and sometimes with ambiguous meaning.

I aim to identify some essential properties which all instances of blame share.[69] In doing so, I will complete the statement of my overarching conception of moral responsibility. I hope to show how this conception has independent merit which recommends it. But setting out this conception also helps to clear the ground for the second part of the thesis, in which I consider one kind of moral responsibility in particular: basic desert. The final section of this chapter is therefore used to introduce basic desert. This is a species of moral responsibility which, despite drawing

[69] My focus will be almost exclusively on blame, and its relation to blameworthiness, rather than praise.

increasing recent attention, remains sorely in need of clarification. I hope, partly through applying the lens provided by my overarching conception of moral responsibility, to help bring some clarity.

This chapter breaks down into four sections. In the first of these, Section 2.2, I set out the wide variety of ways in which blame has been understood in the moral responsibility literature. In Section 2.3 I proceed to offer an account which can make sense of these competing ideas of blame. My account – the *Appraisal Theory* of blame – is one which naturally follows on from and complements my preceding account of agential features and blameworthiness. The result is admittedly a very thin description of what blame is. It claims to describe only a necessary, and not *necessarily* sufficient, requirement: blame necessarily involves an appraisal of those essential constituents of blameworthiness which I outlined in Chapter 1. The thinness of my proposal means it can easily accommodate most existing accounts of blame, although I also discuss what putative kinds of blame it cannot accommodate. Also, my account cannot *explain* what are sometimes thought to be important characteristics of blame; I discuss why I do not view this as a problem for the account. I finish this section by characterising both a moderate and an ambitious version of the account. The ambitious version represents a new way of conceiving blame and praise. The moderate thesis still represents a valuable finding and provides the conceptual clarity needed for my purposes in the rest of the thesis. Section 2.5 summarises my findings across the first two chapters so far in presenting the full overarching conception of moral responsibility, which is now complete. This conception presents a formula for how theories of moral responsibility are constructed. I set this formula out, and then end the chapter by beginning to fill out this formula for basic desert moral responsibility. I briefly outline the core, established tenets of this kind of moral responsibility, and argue that we can understand the agential features implicated in basic desert in terms of their *sheerness*. But I also identify where there are gaps in how basic desert has so far been specified – it is unclear what kind of blame and blameworthiness is at stake. This provides a bridge to Part Two of the thesis, in which I aim to address these gaps, and consider what this means for how we reassess basic desert skepticism.

2.2. Getting A Handle on Blame

I have so far said very little about the kinds of response which the reasons of blameworthiness and praiseworthiness recommend. I have simply referred to these responses generically as *blame* (and *praise*). But any complete conceptualisation of moral responsibility needs to say more about what it understands blame and praise to be. My focus here will be principally on blame, and its relation to blameworthiness, rather than praise. This is mostly for reasons of space (it also reflects where the focus has predominantly been in the literature). However, I generally take my arguments to have extension to praise, too, and will indicate where this is the case. I will therefore now offer an analysis of blame, which forms the third and final component of my overarching conception of moral responsibility.

Some quick initial clarifications. The kind of blame of interest here is different to the sense in which we blame a tsunami for the destruction it wreaks. To blame in that sense means nothing more than to ascribe causal responsibility, usually for something undesirable. Also, whatever blame is, it is typically distinguished from the judgment that one is blameworthy.[70] So, if Fred blames Kevin, we understand this to mean that Fred does something other than forming a judgment that Kevin is blameworthy. This gap between judging blameworthy and blaming is why it seems entirely unproblematic, as Angela Smith notes, to say "I know he is blameworthy for doing X, but I just can't bring myself to blame him for it" (2013: 30). To be clear, a reason of blameworthiness is generated by the existence of certain facts – facts I have specified as including, at minimum, an agent's deployed agential features and subpar moral performance. So, in generating a reason of blameworthiness, these facts thereby render a judgment of the agent's blameworthiness accurate. But the reason itself favours something *distinct* from forming a judgment of blameworthiness. Namely, it favours blame. So, what is this blame?

2.2.1. Blame – Examples and Variables

Blame is a concept I have found especially difficult to pin down, despite its embeddedness both as part of our everyday experience and within theories of moral

[70] Even so, my impression is that the folk do regularly use *I blame X* to mean *I judge X blameworthy* (as well as to mean *I judge X causally responsible*).

responsibility. Others have found this too.[71] Consulting the moral responsibility literature, we encounter a wide array of different theories of blame. Here are some leading examples. Some cognitivist accounts say that we blame others simply when we *judge* that they bear ill will, or that they have a committed a fault (e.g. Hieronymi 2004; Fricker 2016).[72] Other accounts agree that blaming involves judgment, but in combination with something else too. For some this additional element is a *desire* that the target had acted otherwise (e.g. Sher 2006: 93-14; Arpaly and Schroeder 2014: 158-199); for others it is the *modification of our relationship* with those we blame (e.g. Scanlon 2008: 122-214).

One very influential tradition, rooted in P.F. Strawson's (1963) sentimentalist approach, recommends that blame should instead be conceived in *emotional* terms. For Strawson and many of his 'Strawsonian' followers, it is specifically the negative emotion (or reactive attitude) of moral anger which is fundamental to blame.[73] Shoemaker (2015) considers the Strawsonian view to be the leading expression of one type of moral responsibility: accountability. We hold people accountable in response to the quality of their regard, and blame is due in the form of anger where this regard is of poor quality. Shoemaker incorporates accountability within a tri-partite theory of moral responsibility. This theory describes two further kinds of moral responsibility: attributability and answerability. Both, like accountability, describe a distinctive species of emotional blame (and praise), associated respectively with the two other qualities of will I outlined in Section 1.3. So, attributability concerns blame in the form of disdain (and praise in the form of admiration), which responds to quality of

[71] R. Jay Wallace describes blame as "one of those phenomena that is both familiar and philosophically puzzling… A satisfactory analysis is elusive" (2011: 348). Michael McKenna laments how "[d]espite the pervasiveness of the phenomenon in ordinary life, blame is an elusive notion. It is maddeningly hard to nail down a theory that gets the extension even close to right" (2013: 119). Similarly, T.M. Scanlon: "[a]lthough experiences of blame and being blamed are familiar parts of everyday life, it is not entirely clear what blame involves" (2013: 84).

[72] Note, the judgment described in these accounts is distinct from a judgment of *blameworthiness* as discussed in the previous paragraph. Again, a judgment of blameworthiness instead reflects belief in the existence of a normative reason which favours doing something – *blaming*. For such cognitivist accounts, the blame in question happens to take the form of a judgment. But *not* simply a judgment that there is blameworthiness (rather that that there is ill will or moral fault).

[73] Other emotional theories of blame, which treat anger as central to blaming, are offered by Wallace (1994, 2011), Wolf (2011), Bell (2013) and Menges (2017), amongst others.

character.[74] Answerability concerns blame in the form of disapproval (and praise in the form of approval), which responds to quality of judgment.

Each of the above conceptions of blame can be thought of as *constitutive*: they define blame in terms of the content out of which the response is constituted. This content is comprised of mental states (such as judgments and emotions), some of which are conative (such as desires) and motivate certain actions (such as relationship-modification). A large family of views takes a quite different approach, choosing instead to understand blame with reference to a distinctive *function* which it fulfils. Various functions have been understood as essential to blame, including expressing protest (e.g., Talbert 2012; Smith 2013); being a communicative act (Macnamara 2015; Fricker 2016); initiating a conversation with the target (McKenna 2013); signalling commitment to norms which oppose the target's actions (Shoemaker and Vargas 2019); or facilitating shared knowledge of how interpersonal rights and obligations were altered by a wrongdoer's actions (Sliwa 2019). Instrumentalist theories of moral responsibility are not typically grouped with this functionalist family, but I think they reasonably might be. I argued in Chapter 1 that these views should be understood as theorising about something categorically different to non-instrumentalist theories. (In instrumentalist theories, agential features play a fundamentally different and reducible explanatory role in the generation of reasons to blame.) Yet instrumentalist theories are nonetheless in the business of identifying reasons to blame, and the blame they describe can be understood in functionalist terms. On these theories, blame is conceived as a response which, either directly or indirectly, is used as an instrument to fulfil the function of delivering desirable consequences. As already outlined in Chapter 1, direct instrumentalists ascribe this function to discrete occurrences of blame (e.g. Smart 1961; Pereboom 2021a), while indirect instrumentalists ascribe the function to the overall practice within which individual blaming responses occur. Usually, the cited function is to foster moral agency and reasons-responsiveness in both the blamed party and the wider moral community (e.g. McGeer 2015; Jefferson 2019). Finally, it is open to functionalist accounts of blame

[74] The original distinction between accountability and attributability, presented as two different "faces" of moral responsibility, can be credited to Gary Watson (1996). Shoemaker's three-way distinction, which adds answerability into the mix, is one which now seems to be established in the literature and debate.

to be pluralist with regards to blame's content. Different mental states or activities might serve the same function (as noted by Shoemaker & Vargas (2019: 591)).

In addition to these contrasting views about what blame is, there is also some debate about which is the *paradigmatic*, explanatorily basic form of blame, as opposed to forms which are secondary or derivative.[75] Several authors see emotional, angry blame as the central kind. Susan Wolf, for example, claims that "the paradigm of blame involves an "angry" feeling or attitude" (Wolf 2011: 344). Pereboom appears to agree with Wolf, saying "of all the attitudes associated with moral responsibility, it is anger that seems most closely connected with it" (2001: 208). Fricker (2016), alternatively, claims it is communicative blame which is paradigmatic.

Even once we have settled its content or function, and any question of paradigmatic status, there are further complexities or *variables* attending blame. It is also important, for the purposes of clarity, to have these on the table. One important variable concerns where the blamer is situated in relation to the blamed agent. The literature tends to agree on three possibilities here. First, there is self-blame, where we blame ourselves. Then there are two interpersonal modes of blame. Second-personal blame, where we blame one whose moral underperformance has personally wronged us; and third-personal blame where the blamer is not themselves a victim of the wrong in question. In some cases, distinct concepts have emerged to demarcate these respective blaming relationships.

The clearest example is perhaps those Strawsonian reactive attitudes which figure in accountability (as discussed above). These emotions of moral anger are standardly termed *guilt* where self-directed, *resentment* if directed towards another who has

[75] At least, this is a debate amongst those who hold a pluralist view of blame, one which allows that there is more than one kind of blame. Whereas for those authors who deem only one kind of response to be properly describable as blame, presumably that response will automatically gain paradigm status. Emotional theorists of blame in particular have often held this monistic view. Strawson (1963: 208, original italics) for instance claims that the emotional reactive attitudes alone can tell us "*all* we mean, when… we speak of desert, responsibility, guilt" (and also condemnation and justice, to boot). Similarly, Wallace suggests it is "strange to suppose that one might blame another person without feeling an attitude of indignation or resentment toward the person" (1994: 75). Bell (2013) and Shoemaker (2015) take a similar view.

shown you poor quality of regard, and *indignation* where their disregard is shown to someone else.[76]

Another variable, pertaining to at least some proposed versions of blame, concerns whether blame is expressed publicly, to others, or is kept private. For instance, indignation can be inwardly felt whilst we remain unmoved on the surface. It can also be externally expressed in violent outrage, or in barely discernible changes in our facial expression or behaviour. A further, final variable concerns the strength of blame. Blame often come in degrees. The same kind of blame – whether a reactive attitude, or a protest or norm-reinforcing signal, or the modification of one's relationship with blame's target, say – can be experienced or expressed more or less intensely.[77] Here a widely accepted proportionality constraint is relevant. It is generally understood that the degree of blame one receives ought, all-else-being-equal, to be commensurate with the degree of blameworthiness. (Where, according to my analysis, the degree of blameworthiness is at least partly a function of the severity of the moral underperformance, and the robustness of the agential features deployed – the amount of control and knowledge they endow the agent.) Whether one's blame is proportional may also be influenced by further facts which do *not* affect the target's blameworthiness. For instance, facts about the blamer's hypocrisy or complicity (as discussed in the previous chapter), or about whether the wrongdoer has reproached themselves or offered redress (as discussed in Smith 2007).

All of this makes for a very diverse range of proposals for what blame is, and how it might manifest depending on how the additional variables are realised. The foregoing is not intended as a comprehensive survey of how blame has been understood in the moral responsibility debate.[78] A full survey is beyond my scope, but nor is it required given my purposes. I will not offer any decisive verdict about which kinds of response properly qualify as blame, less still which of them are paradigmatic. For now, I am

[76] This terminology originates with Strawson himself (1963: 190-207). Many have since taken it up (see Wallace 1994: 51-52; Fischer & Ravizza 1998: 5-8; McKenna 2012: 71-73).

[77] These variables further complicate the debate about which kind of blame is paradigmatic. McKenna, for instance, sees what he calls "directed blame" – second-personal blame which is overtly expressed towards the blamed party – as the "more fundamental notion" (2013: 121). Others instead claim that self-blame is explanatorily basic, usually in the form of guilt (for example see Carlsson 2017 and Duggan 2018).

[78] Broadly similar taxonomies of blame are offered by Coates & Tognazzini 2012, Smith 2013, and Tognazzini & Coates 2021.

simply aiming to register the sheer variety in how blaming (and praising) responses have been conceived, and the variables which complicate the picture even further.[79] (Or, translating this into my preferred terms, how the deployment of agential features in negative moral performance has been thought to generate normative reasons in favour of a wide variety of responses). This already serves to underline the potential problem in talking generically – as I have myself, prior to this section – of blame and praise. These generic concepts are inexact and open to variable interpretation. A guiding aim of this thesis is to avoid inexactness of this sort (one which has afflicted the debate around skepticism about basic desert, as I discuss further in this chapter's final section).

Given such divergence in how blame has been understood, what unity might be brought to bear? With this mind I now offer my conception of blame.

2.3. The *Appraisal Theory* of Blame: A Thin Conception

My proposal for how we should conceptualise blame is already implied by my analysis of blameworthiness. Put concisely, I propose we understand blame to be a response which results from an appraisal of the two essential constituents of blameworthiness. According to the account of blameworthiness offered in the previous chapter, this means a blamer necessarily appraises the following two facts about the target of their blame. First, that they have morally underperformed. Second, that they possess agential features which were deployed in that moral underperformance.[80] I have argued that any instance of blameworthiness is grounded, at least partly, in these two facts. So, for any response to qualify as blame, it must involve an appraisal of these essential constituents of blameworthiness.[81] I will refer to this qualifying requirement

[79] Very plausibly this variety and complexity extends to praise too. Standalone accounts of praise have been relatively rare (though some have emerged more recently. Telech's (2021) conversational account and Holroyd's (2023) social function account are both insightful views of praise). But many of the constitutive and functional accounts of blame I have mentioned recommend (whether implicitly or explicitly) how we might conceive of praise in analogue fashion. Like most of these authors, however, I will remain principally focused on analysing blame. (In Chapter 4 I return to the issue of why praise has received less attention).

[80] By *agential features* I here mean to include both epistemic and (what I have termed) control-conferring agential features, as discussed in Chapter 1.

[81] Crucially, this does *not* mean that the appraised facts necessarily reflect genuine features of the target which do in fact generate a reason of blameworthiness. Rather, what is needed for blame is merely the appraisal of the two *kinds* of fact which (I argue) are the essential constituents for generating any reason

as the *Appraisal Requirement*. It is what defines my *Appraisal Theory* of blame. (My analysis of praiseworthiness implies a symmetrical qualifying requirement for praise. The only difference being that the requisite appraisal is of morally *successful* performance, in which agential features are deployed. So, with this modification, my conception of blame might be extended to praise too. But I leave any further exploration of this possibility for future work).

By an *appraisal* of the essential constituents of blameworthiness I mean an evaluative representation of those facts in a very inclusive sense. This *may* take the form of a judgment or belief that those facts obtain, but it need not rise to the level of settled judgment. It might instead be a non-judgmental representation. Here I have in mind the kind of appraisal described in certain theories of emotion (including emotional blame). For instance, here is how D'Arms & Jacobson talk about the emotion of shame:

> A bout of shame presents its object (typically some feature of the agent) as shameful; but to be ashamed is not necessarily to endorse this presentation [of shamefulness] in any way, much less in the particular manner which entails judging that this feature is genuinely shameful (2000: 67).

They describe the kind of representation in question as "quasi-judgmental, quasi-perceptual" (2000: 68).[82] What goes for shame, I propose, goes for blaming responses in general. To blame is to present or appraise the object of blame as exhibiting the *essential constituents of blameworthiness*. This appraisal might take the form of a judgment, or it might take a different form.

Shoemaker (2015) utilises this notion of quasi-judgmental appraisal in his account of the blaming reactive attitudes. As discussed, for Shoemaker, it is our appraisal of various qualities of will – regard, character or judgment – which gives rise to distinctive blaming emotions. But the appraisal and the ensuing emotion are "encapsulated from judgment" (2015: 22). This means that not only does the appraisal

---

of blameworthiness. Whether the target *is* in fact blameworthy is a different matter entirely, concerning the *existence* (rather than merely an appraisal) of the essential constituents of blameworthiness. I return to this point in Section 2.3.2 in my discussion of blaming the non-blameworthy.

[82] Similarly, D'Arms & Jacobson elsewhere discuss how some emotions (such as shame) are not "cognitively sharpened", as in we need not have formed any particular judgment in order to experience them (2003: 143).

not entail forming any judgment, but that one can even judge that the relevant agential feature is *lacking* and still experience the blaming emotion. For example, my considered judgment might be that your uninterested reply did not betray any disregard towards me. Yet I might still experience resentment towards you, so long as I appraise you as disregarding me in some other, non-judgmental (or 'quasi-judgmental') sense. In such cases the blame persists, "stably recalcitrant" in the face of the contrary judgment (Shoemaker 2015: 91). In the same vein, Pickard (2013) develops the idea of "irrational blame" where we blame despite holding a contrary judgment about the target's lack of blameworthiness. Pickard, like Shoemaker, is offering an account of strictly emotional blame. But perhaps responses other than emotions can also be prompted by such quasi-judgmental appraisals, and thus exhibit recalcitrance in the same way. I will not settle this here. I only insist that if there *are* such responses, then if they are elicited by the responder's appraisal of both essential constituents of blameworthiness, then they too satisfy the *Appraisal Requirement* and qualify as blame. This is true even if the agent simultaneously judges (or otherwise appraises) contrarily that one or both essential constituents are lacking.

Importantly, the sort of appraisal I describe need not be understood by the appraising *blamer* in terms of agential features or blameworthiness. If the folk engage in blame – and presumably any plausible conception of blame will allow that they often do – then surely it is without them needing to have a clear grasp of the agential features possessed by the object of their blame. Nor whether these agential features are control-conferring or epistemic (in the senses I have stipulated), or both. Nor whether these features are comprised out of mental states or events, or otherwise. Nor what species of control (sourcehood or ability to do otherwise, say) they endow. Less still an ability (or inclination) to articulate any of these phenomena in such terms. The *Appraisal Theory* of blame does not impose any of these demands on would-be blamers. It only demands that they appraise the target as having some notion, however crude, of *control* over their moral underperformance. If they do this, they have appraised the two essential constituents of blameworthiness. So long as this appraisal then explains their response to the agent, then my account deems that response to qualify as blame. This notion of control admittedly remains very loose – but I think this reflects the looseness with which it generally features in everyday blaming.[83] The need to tighten up the concept arises when it comes to investigating blame*worthiness*: is it possible to appraise control *of a sort that genuinely generates (in tandem with disappointing moral performance) a reason to respond in a particular way?* This is where, as per the arguments of Chapter 1, I recommend we analyse the relevant control in terms of deployed agential features. If any such control in fact exists, it will be explicable in these terms.

2.3.1. A View with a Lot of Room

The *Appraisal Theory* amounts to an extremely thin conception of blame (and, by extension, praise). It adds, in its *Appraisal Requirement*, little that was not already promised by my account of blameworthiness. It thus sets a very low qualifying standard for blame – one which is entirely agnostic with regards to the content (beyond the requisite appraisal) and function of blame. A merit of my view, as I see it, is that it thereby readily accommodates – or at least it is reconcilable to – most accounts of blame advanced in the literature, including those discussed in the previous section. The kinds of constitutive and functional blame described in these accounts typically will be, and certainly *can* be, issued in response to an appraisal of agential control and moral underperformance. Wherever they are, they satisfy the *Appraisal Requirement*, and my account concurs that such responses qualify as blame.

In fact, the only view of blame which looks obviously difficult to reconcile with my account is Fricker's (2016) account. As discussed, this account says that blame can be comprised of a judgment. Fricker describes the relevant judgment as a "finding fault with someone for their (inward or outward) conduct" (2016: 170). My reading of Fricker's account is that the relevant judgment does not necessitate any appraisal of agential control, merely moral underperformance.[84] In which case, contra Fricker, the *Appraisal Theory* does not consider this judgment sufficient for blame. It fails to

[83] So, my view places no limits on the kind of agential control which must be appraised in order to blame. Control of any flavour will do, however weak or strong, and however far-fetched. Suppose I ascribe supreme powers of control to you – the strongest conceivable form of agent-causal libertarian freedom, say. If I then appraise that you deployed these powers in moral underperformance, and my response to you is prompted by this appraisal, then *whatever* the nature of that response, it is blame. The *Appraisal Requirement* is met.

[84] Which seems a coherent proposal. We can feasibly commit faults – violate moral norms – whilst having no agential control over them (for instance, if our committing the fault was entirely a product of luck).

satisfy one of the two components of the *Appraisal Requirement*.[85] (Still, even this judgment is one which in principle could be, and in practice surely often will be, supplemented with and generated by an appraisal of agential control. Suppose I judge that you are at fault *because* you deployed an agential feature – your quality of regard, say – in a particular way. In this case the judgment *would* then satisfy the *Appraisal Requirement* and qualify as blame). Hieronymi's judgment-based view is different. On her view the relevant judgment is a judgment of an agent's ill will which, channelling Strawson, means that they "failed to show proper regard for others" (2004: 115). As discussed in Section 1.3, my expansive understanding of agential control interprets quality of regard as a form of control. So, this judgment automatically satisfies the *Appraisal Requirement*. It is an appraisal of agential features deployed in negative moral performance.

Indeed, my conception of blame is perhaps most closely affiliated with the Hieronymi-style cognitivist view. That view understands blame to be a judgment about the existence of certain facts. My *Appraisal Theory* likewise sets a qualifying condition on blame in terms of the evaluative content which it appraises, rather than in terms of the psychological states or dispositions that *constitute* the ensuing response. But my account also borrows from emotional theories of blame, in a sense – it widens the range of relevant appraisals so that these include more than just judgments. Also, unlike these cognitivist views, the *Appraisal Theory* does not say that blame is identified with and consists only in the appraisal or judgment. Rather, on my view, blame *also* refers to whatever response ensues due to the appraisal (whether it is an emotion, relationship-modification, a response with a particular function, or anything else). This reiterates how the *Appraisal Requirement* provides only a qualifying condition for blame; I do not claim it provides a complete description of what blame is. I return to this point below.

My account, in virtue of its thinness, also has a relatively easy time accommodating what are sometimes seen as tricky-to-accommodate cases of blame. Shoemaker and Vargas describe several such cases – or "data points" – taken from our blaming

[85] We might think that similar considerations apply to accounts like Sher's or Scanlon's which also understand blame in terms of a judgment, plus (for Sher) a desire or (for Scanlon) a modification of a relationship. But these accounts are explicit (in contrast to my reading of Fricker) that blame involves a judgment about (poor) qualities of the agent's character or judgment, respectively. In which case, the blame they describe necessarily satisfies the *Appraisal Requirement*.

practices. These include blaming oneself, blaming the dead, blaming without any emotion, and blaming hypocritically (2019: 582-53). They argue that various leading accounts of blame, both constitutive and functional, struggle to account for one or more of these cases. They each undertake "fancy dancing" – *ad hoc* revisions to their stipulated conditions of blame – in unconvincing attempts to accommodate them (2019: 582). (In contrast Shoemaker and Vargas claim that their own functionalist theory of blame – the signalling account – can satisfactorily, and uniquely, accommodate the various cases). But whatever the difficulties faced by those other accounts, my *Appraisal Theory* faces no such problems. It seems clear that we can engage in self-blame, blame of the dead, dispassionate and hypocritical blame whilst in each case having appraised the targets of our blame to have exercised control over their moral underperformance. (Perhaps simultaneously. Imagine that Donald, a prolific liar, emotionlessly shakes his head when he thinks of the lies his deceased father used to tell. "He really should have known better than to lie. As should I now".) So, each of these cases of blame is compatible with satisfying the *Appraisal Requirement*, and thus can be accounted for as blame by my view.[86]

I should also note that the *Appraisal Theory* is so inclusive that a judgment of blameworthiness might qualify as blame. Indeed, given my account of blameworthiness, this is a response we should *expect* to naturally follow an appraisal of agential features deployed in moral underperformance (at least if it hopes to be an *accurate* judgment). I said at the outset of Section 2.2 that the kind of blame which

[86] There is another of Shoemaker's and Vargas' data points of blame which it seems even my thin conception cannot accommodate. Albeit one which is rarely discussed – in fact I have only seen it described by them. They call it "hypothetical blame" (2019: 595). These are cases where we (purportedly) blame another even though we do not take them to have committed any moral infringement. The example they discuss is one where, if we are watching a film and we see a character behave badly, I blame you because I predict you would behave in the same way as the character. Shoemaker and Vargas recognise that one might reasonably question whether this really *is* blame. Certainly, I would. It explicitly flouts the *Appraisal Requirement*: it is a response which involves no appraisal of moral underperformance in the blamed party. (It thereby also necessarily fails to satisfy the second element of the requisite appraisal: that agential features are deployed in the carrying out of said moral underperformance. If there is no underperformance, there can be no such deployment). That said, ultimately, I would agree with Shoemaker's and Vargas' sentiment that "whether or not this is a case of blame (as it is presently understood) is mostly beside the point" (2019: 595). They deem it beside the point because their principal aim is to articulate a family of responses which share a particular and important function, rather than to conclusively show this family of responses is co-extensive with blame. Likewise, my aim is not to decisively establish sufficient conditions for blame (although in Section 2.3.4 I do discuss a more ambitious version of my account which interprets the *Appraisal Requirement* as a sufficient condition). Rather, my principal aim is to identify an important (and I think underappreciated) qualifying requirement, whose identification also provides much-needed conceptual clarity ahead of my analysis of basic desert skepticism.

theorists of blame are interested in articulating should be distinguished from the judgment of blameworthiness. This remains true. Even though judgments of blameworthiness can qualify as blame on the *Appraisal Theory*, those judgments are judgments about the existence of a normative reason favouring something *distinct* from a judgment of blameworthiness. This other, distinct thing is the blame we are seeking to conceptualise. It is just that my conception of *that* blame happens to be one which accommodates judgments of blameworthiness too.

2.3.2. Some Implications of the *Appraisal Theory*

Automatic and Non-Automatic Satisfaction of the *Appraisal Requirement*

I have said that my account accommodates most existing views of blame, since those views describe responses which *can* result from an appraisal of the essential constituents of blameworthiness. Some of these responses necessarily *entail* the requisite appraisal (such as the blaming reactive attitudes in the Strawson-Shoemaker sense, and the blaming judgment described by Hieronymi. As discussed, both entail an appraisal of poor quality of will). These therefore automatically meet the *Appraisal Requirement* and qualify as blame. But other notions of blame do not come with the requisite appraisal embedded in this way. This seems most clearly true of functional blame. Various functions which have been associated with blame might feasibly be served *without* it necessarily needing to be the case that the blamed party is appraised as having morally underperformed whilst exercising control. For instance, consider some of the blame functions we have already mentioned: expressing protest; signalling one's commitments to moral norms; instrumentally fostering moral agency within a moral community. It seems very possible that blame might realise such functions even in cases where the blamer does *not* appraise the target as having deployed their own agential features in moral underperformance. Here lie familiar worries about *instrumentalisation*: if blame is understood in purely functional terms, this may sometimes recommend blaming agents who are innocent of any wrongdoing so as to achieve the functional aims (Jefferson 2019: 557).

Functional accounts of blame have typically sought to avoid such unsavoury implications by adding constraints on the blame they describe (and defend). One approach is to set extra, non-functional conditions on blameworthiness which must be satisfied for blame to be warranted. For example, McKenna (2012) understands blame in terms of its functional role as part of a reparative conversation. But he also insists that blameworthiness requires that the blamed party exercised reasons-responsiveness over that for which they are blamed. Other authors see blame's function as registering moral protest, but separately specify that blameworthiness entails exhibiting quality of will in the sense required for answerability (Smith 2013) or attributability (Talbert 2022). Some indirect instrumentalist accounts argue that the instrumental function of blame is best fulfilled by a practice where only those agents appraised as having the relevant agential features for blameworthiness are blamed (Jefferson (2019); McGeer (2015); McGeer & Pettit (2015)).[87] Alternatively, these accounts might build in non-consequentialist side constraints. Jefferson (2019: 570) for instance suggests that considerations of fairness and rights can ensure that functional blame is targeted only at the blameworthy.

I will not assess the success of these various strategies. But I suggest we can effectively view them as attempts to reconcile these accounts of functional blame with the *Appraisal Requirement*. (Or, equivalently, to ensure the functional blame they endorse is *genuinely* blame, on the terms of my account). It is because purely functional accounts of blame do *not* automatically satisfy the *Appraisal Requirement* that they must adopt such strategies.

Cynical Blame is Not Genuine Blame

If satisfying the *Appraisal Requirement* is necessary for blame, this entails that we cannot blame without making an appraisal in which both essential constituents of blameworthiness are present (even if we simultaneously make one or more other appraisals in which one or both is not present).[88] This is typically how we appraise inanimate entities – a wall, or a cloud, for instance. The same with new-born babies or animals. None of these are the sorts of thing we appraise as having morally underperformed, never mind as having exercised control in doing so. Such things are

[87] The indirect instrumentalist accounts mentioned, however, understand an agent's blameworthiness in terms of their general capacity to be *influenced* by blame. This in turn requires the agent to be reasons-responsive or sensitive to the moral considerations which blame communicates (Jefferson 2019: 559; McGeer 2015: 2642). It is not clear to me that such blameworthiness does in fact necessitate that the agent deployed any particular agential feature at the (earlier) time of the action for which they are blamed. In which case, appraising such blameworthiness would *not* necessarily ensure the *Appraisal Requirement* is satisfied.

[88] The parenthesis describes those cases of recalcitrance discussed in Section 2.3, where we satisfy the *Appraisal Requirement* (and thus we blame) whilst simultaneously making a contrary appraisal that one or both of the essential constituents of blameworthiness are *lacking*.

generally not taken to be *blameable* (in the sense relevant to moral responsibility). My account explains why this might be.

Similarly, and perhaps less intuitively, my account implies that *cynical* blame is not really blame at all. By cynical blame I mean blame directed at agents who are appraised as *not* having deployed their agential features in moral underperformance (or at least not in a way which is commensurate with the blame). A familiar kind of cynical blame is scapegoating. An innocent scapegoat is precisely an agent who is 'blamed' by a blamer who appraises them as non-blameworthy.[89] Scapegoating therefore expressly fails to meet the *Appraisal Requirement* and does not qualify as blame, on my account. It is a different category of response (perhaps we might think of it as inauthentic or pretend blame). Another form of cynical blame is 'blaming' in the knowledge that the target of blame has a valid excuse for their wrongdoing. The blamer appraises the target as having morally underperformed, but not due to their deployed agential features (so the target is appraised as satisfying only one of the two essential constituents of blameworthiness). Again, the *Appraisal Requirement* is unmet. This too is not *really* blame.

With cynical blame, it always seems possible to say, *sotto voce*, 'of course, I scolded Bill, but I do not *really* think Bill is blameworthy at all. It is Ben that I secretly resent for what happened'. So, the *Appraisal Theory* adjudges it is Ben, and not Bill, who is receiving blame here. Again, what determines whether a response is blame is not its constitutive mental states or activities, but the content of the appraisal which provokes these things. This suggests a further implication of my view: blame may often be difficult to identify. Two responses might be identical in terms of their constitution or function, but it may be impossible to *tell* whether one or both are blame. For instance, suppose Martha modifies her relationship with Bill and Ben in exactly the same way,

[89] Though scapegoats, it seems, *will* usually be appraised by the blamer as possessing the kind of agential features and control which would make them *capable* of the blameworthiness of which they are accused. This ensures that third parties might be convinced (mistakenly) of their blameworthiness. Whereas agents who lack agential features, and are thus *exempted* from moral responsibility, are not normally picked as scapegoats. (Hence why a child would not (sensibly) be scapegoated for a crime which only a criminal mastermind could pull off. Not that I endorse or wish to offer advice on effective scapegoating!). Crucially though, the scapegoat is not appraised by the scapegoater as having deployed their agential features in the wrongdoing for which they are blamed (the scapegoater knows the scapegoat has *not* committed the wrong in question). This is why, according to the *Appraisal Theory,* they are not *genuinely* being blamed by the scapegoater. Whereas third parties *may* yet be genuinely blaming them, if they are duped into mistakenly appraising the scapegoater as blameworthy.

shunning both and withdrawing her friendship. If we later learn that it is only Ben who she treated this way due to appraising in him the essential constituents of blameworthiness, it is only then that we confirm that she was blaming Ben, and not Bill.

The Possibility of Blaming the Non-Blameworthy

One point is worth reiterating here, as my terminology may be liable to mislead. A desideratum for any account of blame is that it makes room for the possibility of blaming the non-blameworthy. Blame surely often misses it target, as does praise. It would be absurd to propose that those who receive blame are always blameworthy. The *Appraisal Theory*, however, forges an intimate connection between the conditions of blame and blameworthiness. It says that if we are to blame an agent, we must appraise in them the essential constituents of blameworthiness. Does this still leave room for blaming the *non*-blameworthy?

Yes, plenty. When we appraise the essential constituents of blameworthiness, this does *not* guarantee that the agent we appraise is in fact blameworthy. What is appraised may not reflect the facts as they are – the appraisal might be erroneous in three possible ways. First, we could be mistaken that the agent in fact morally underperformed.[90] (Perhaps Jerome's apparently insulting comment was misinterpreted). Second, we might be correct that the agent morally underperformed, but mistaken that they exercised any control over this. (Perhaps Jerome's comment *was* insulting, but it was involuntary and not endorsed by his real self). Third, our appraisal might be correct about the deployment of agential features in moral underperformance, but these facts may be insufficient to generate blameworthiness. (Perhaps Jerome's insult *was* voluntary and endorsed by his real self, but neither of these are sufficient to generate

[90] This kind of error also bears upon the fairly common practice of blaming young children for their misbehaviour. This blame is often functional in nature, intended to reinforce moral norms and promote improved future behaviour (praise is used in the same way). Young children are usually considered exempt from moral responsibility and blameworthiness, however. (See Fischer & Ravizza 1998: 76–81; McKenna 2018: 43; Shoemaker 2015: 14; Jefferson 2019: 568 for discussion of the exempted status of young children.) So, this is an example of 'blaming' the non-blameworthy. Young children plausibly *do* exert some degree of control – voluntariness, at minimum – over the actions for which they are blamed. But they are typically not seen, at least when *very* young, as subject to the demands of morality, meaning their bad behaviour does not register as a moral underperformance. Assuming they are appraised in such terms when they are subjected to functional 'blame', then they are appraised as exhibiting at most only one of the two essential constituents of blameworthiness (exercise of control). So, the *Appraisal Theory* would not deem this kind of functional blaming as genuine blame.

blameworthiness).[91] The first two errors are empirical; the last error is better described as ethical or rational.

So, blameworthiness is a matter of whether particular facts are *true* of an agent, and whether those facts generate a reason in favour of responding in a distinctive way. Whereas identifying blame is a question of whether a response is had upon appraising *those kinds of facts*. The only relevant information for establishing whether a response qualifies as blame, then, according to the *Appraisal Theory*, are facts about the responder – specifically about what they appraise. Where the *Appraisal Requirement* is satisfied, and we blame, this tells us nothing about whether that blame is supported by reasons of blameworthiness. Indeed, on my view, blaming would remain possible in a world in which no one possesses agential features, and thus where there is no blameworthiness whatsoever. We could still appraise the *essential constituents* of blameworthiness – deployed agential features and moral underperformance – in such a world. It is just that the appraisal would always misrepresent reality.

2.3.3. Some Worries for the *Appraisal Theory*

Too Inclusive?

One might worry that the generous accommodativeness of my conception of blame also has significant costs. The bar set by the *Appraisal Requirement* is so low that in principle *any* response might qualify, provided it is a response to the requisite things. This delivers counter-intuitive results in terms of what it classifies as blame, and so perhaps it is *over*-inclusive.

For instance, suppose Joni is appraised by Nina as having stolen from Aretha, and having done so whilst being fully responsive to reasons. Nina deems theft to be a serious moral wrong. Joni is thereby appraised as having satisfied the essential constituents of blameworthiness: of having failed to meet moral expectations, with this failure attributed to her deployed agential features. (Again, Nina might simply experience this as an appraisal of some notion of Joni's control over her actions). If

[91] Similarly, the *Appraisal Theory* also allows for *disproportionate* blame. For instance, perhaps Jerome's insult was a minor *faux pas*, but it is met with a vituperative expression of indignation. Assume that Jerome is blameworthy – his *faux pas* does generate a reason in favour of *some* response (a very mild rebuke). If the vituperatively indignant response issues from an appraisal of the essential constituents of blameworthiness in Jerome, then my account deems that response to be blame, regardless of how disproportionate it is to Jerome's actual blameworthiness.

Nina chooses to somersault due to her appraisal of Joni, then the *Appraisal Theory* deems her somersault to be an instance of blame. Equally if Nina responds with a gleeful smile, this too would be understood as a blaming response according to my account. The *Appraisal Requirement* remains met. We would of course find such responses extremely odd. Such eccentric or positive responses would not be recognisable as blame. We might find it difficult to comprehend that Nina really *did* appraise Joni has having morally underperformed, at least in a way which Nina takes seriously.

The fact that the *Appraisal Theory* affirms these responses as blame clearly stretches our ordinary understanding of what blame is. This may seem an unpalatable implication of the view. However, I do not think this should necessarily be viewed as a mark against the theory. True, these are not cases of blame we are ever likely to come across in our everyday moral responsibility practices. But the best explanation of this, I think, is *not* that they are not blame. Rather, it is that they are not the *kind* of blame which our appraisals of agential features, deployed in moral underperformance, ever in fact tend to motivate. I will build on this line of reply in Section 2.3.4, where I suggest how my account suggests a new approach for thinking about blame.

No Explanation of Blame's Familiar Qualities

Another possible objection to my account is that it fails to explain what are sometimes taken to be essential, or at least frequently exhibited, qualities of blame. Blame has been described as having a "force" or "depth" which the mere ascription of causal responsibility lacks (Wolf 1990; Hieronymi 2004). More concretely, blame is also often understood to have a characteristic *sting* (Pickard 2013). It is typically unpleasant to receive, whether from others (say, as resentment) or from ourselves (say, as guilt). Relatedly, blame is sometimes observed to have a propensity to *harm*. McKenna, for instance, identifies three distinct harms which blame inflicts upon the "welfare interests" of the blamed party. It negatively affects their friendships; it involves interference from others in their personal life; and it can undermine their emotional stability. He thinks these are often "very mild" harms, but harmful nonetheless (McKenna 2012: 138-141; also 2024: 36).[92] Some argue that the negative reactive

[92] Bennett (2002: 148-151), focusing on the negative reactive attitudes, gives a similar account of blame's harmfulness.

attitudes are inherently harmful. (For instance, Pereboom (2021b) has recently discussed at length the potential for interpersonal anger and resentment to inflict harm. Carlsson writes that "to blame oneself – to feel guilty – necessarily involves suffering" (2017: 90).) Blame's potential for harm therefore raises the stakes; it means blaming the non-blameworthy can be a cause for serious moral concern.

The *Appraisal Theory* – and its *Appraisal Requirement* – does not seem to offer any ready explanation of blame's forcefulness, or its tendency to sting or harm.[93] Nor indeed does it offer a ready explanation of any *further* qualities of blame which might be posited. Again, the *Appraisal Theory* is entirely agnostic with regards to the constitutive and functional properties of blame. If a response satisfies the *Appraisal Requirement*, this tells us only what it is a response *to* – it does not specify *any* further qualities of the response. So, the connection to blame's distinctive force, sting and harm, for instance, seems lost. Is this a shortcoming of the view?

In reply, I first note that this is not a worry unique to the *Appraisal Theory*. For example, views which identify blame simply with a judgment, or with a judgment in combination with a desire, face difficulty in accounting for stinging or harmful blame. The *Appraisal Theory* can at least account for blame which has such qualities (as such blame can in principle satisfy the *Appraisal Requirement*). It just cannot obviously *explain* why such qualities are especially prevalent.

Still, the *Appraisal Requirement* at least *begins*, I think, to explain why blame might typically sting or harm. It says that blame necessarily involves a twofold disappointment in the eyes of the blamer. First, that you failed to meet a moral standard which they take seriously. Second, compounding this, that you were not helpless or passive in doing so. You had some kind of control over this failure. To be appraised in such a disappointed fashion *already* hurts, and gives blame something of a sting.[94]

[93] In parallel, praising responses are widely understood as favourable for the recipient. Gratitude, admiration and suchlike are typically described as pleasurable responses to receive, and not thought to cause harm (as noted by Bingeman 2024). If we assume the *Appraisal Theory* extends to praise, the theory cannot obviously explain these characteristic qualitative properties either.

[94] Is not blame hurtful in this sense only if we assume the blamed person shares the moral standards of the blamer? If the target of blame instead endorses *opposing* standards, then blame – at least certain kinds of it – might well seem a *desirable* thing to receive (and praise undesirable). Receiving blame from one we take to have morally perverted values – imagine they believe that wanton killing is morally virtuous and blame us for failing to engage in it – would be evidence that one has got something right. We will have deployed our agential capacities towards moral *success*, on our terms. Their judgment of

Also, if appraised facts of this kind are capable of generating any reason in favour of a distinctive kind of response, we might think it *likely* the response would be one which is unfavourable, and perhaps even harmful, to the agent. This is why somersaults and gleeful smiles seem out of place in response to conduct which we appraise as morally subpar – these are not responses which seem to *fit* such appraisals. Still, beyond these somewhat weak mitigations, I accept that the *Appraisal Theory* itself provides no further explanation, and certainly no guarantee, of blame's typical sting or harmfulness. Again, however, I do not think this is necessarily a flaw of the account. As I will discuss in the next section, I am not convinced that the onus should be on a theory of blame to explain the qualitative features of the blame we in fact experience.

2.3.4. A New Way of Thinking About Blame?

Both potential objections raised in the previous section concern how the *Appraisal Theory* delivers verdicts about blame which are at odds with our blaming practices and experience. The theory seems to classify too much as blame, and it fails to explain why blame has the qualities we typically associate with it. I will close my analysis of blame by sketching a moderate and a more ambitious version of the theory, each of which provides a way of defusing these kinds of objection.

The Ambitious Thesis

The more ambitious thesis says that the *Appraisal Requirement* captures *all* that blame is. It represents a *sufficient* requirement and not merely one qualifying requirement amongst others. This does not mean that satisfying the *Appraisal Requirement* exhausts the description of what blame is. As we have seen, many different blame proposals – with a diverse range of constitutive and functional descriptive properties – can satisfy it. But, according to the ambitious thesis, the *Appraisal Requirement* provides the only condition which must be met for a response to qualify as blame.

What motivates this ambitious thesis are my doubts about how instructive the data points from our actual blaming practices really are in setting the conceptual boundaries on blame. Take the observation that blame often stings or is harmful. This observation

our blameworthiness – or at least, their appraisal of our having exercised some control over the relevant (in)actions – would in this case be welcome, not hurtful. Even so, the response which their appraisal motivates (such as their resentment, or admonishment, or punishment) might nonetheless remain stinging and harmful.

reflects how our appraisals of agential features deployed in disappointing moral performance – what I have called the essential constituents of blameworthiness – overwhelmingly tend to elicit responses with these distinctive qualities. This is true whether these responses are elicited instinctively (as with the emotional responses of guilt or resentment) or more deliberatively (as with protest or the cooling of a friendship, say). Whereas, somersaults and gleeful smiles are not part of our blaming repertoire just because these same appraisals do *not* tend to elicit such responses. So, blame's tendency to sting and to harm, and to not take eccentric forms such as somersaults, are in an important sense *contingent* qualities. Rather than being conceptually inherent to blame, they are contingent upon how agents do in fact respond to the relevant appraisals.[95]

Seen this way, blame which has these familiar but contingent qualities comprises only a *subset* of all (conceptually) possible blaming responses. If we refer only to this subset of responses in determining the conceptual boundaries of blame, we risk *biasing* our account towards the contingent features of this subset. It of course remains sensible to prefer an account of blame which can *accommodate* blame as we experience it. (An account which ignored our blaming practices and did not recognise most of what we consider to be blame as genuine blame would seem implausible). The *Appraisal Theory* achieves this accommodation: it just accommodates much more besides.[96] But to *delimit* the concept of blame in terms of these contingent qualities or data points is, I claim, misguided.

It also remains enormously important that we interrogate this particular subset of blame, since this is the blame *in which we in fact engage*. We should hope that this blame is tethered, proportionately, to genuine reasons of blameworthiness, so that we are blaming for good reasons. Otherwise, we are blaming the non-blameworthy and potentially inflicting unwarranted harm. This will depend, in the first instance, on whether our appraisals of agential features, deployed in moral underperformance, are accurate. Then, if so, what set of responses do these facts generate reasons in favour

[95] By extension, the same goes for the tendency of praise to be experienced pleasurably and harmlessly. These too are qualities which are contingent upon the actual responses we have to the appraisal of deployed agential features in positive moral performance.

[96] Well, with some exceptions, such as cynical (so-called) blame and certain kinds of functional blame (such as 'blaming' young children).

of? Is our stinging and harmful blame amongst this set, or does our blaming misinterpret the normative significance of these facts? These however are questions for theories of *blameworthiness* to investigate, in conversation with those who are *skeptical* about moral responsibility and blameworthiness. Part Two of this thesis is where I aim to make my own contribution to this conversation.

The ambitious version of the *Appraisal Theory* therefore represents, I think, a different approach to thinking about blame. It insists that a theory of blame is not obliged to explain the qualities and the narrowness of our actual blaming repertoire. These features are instead explained by our contingent patterns of blaming, which in turn reflect how we in fact react to appraisals of agential features deployed in moral underperformance. The worries of the previous section, concerning the *Appraisal Theory*'s over-inclusiveness and lack of explanatory power, are therefore defused. They each reflect standards which we should not expect a theory of blame to meet.

The Moderate Thesis

I appreciate that the ambitious thesis may leave some unconvinced. There is little consensus about what blame is, and I do not expect to achieve consensus here. Perhaps the *Appraisal Theory* does cast the net too widely and accommodates responses which, in the final analysis, should not properly be categorised as blame. We may want, in a sense, *more* from a theory of blame, so that it more closely tracks and explains our actual blaming practices. This depends on what we want our theory of blame to achieve, a question which I ultimately leave open. Nor do I wish to be territorial over this conceptual terrain – perhaps more than one theory of blame is viable. Indeed, it is possible that no unified set of sufficient conditions *can* account for the diversity of our blaming practices in a way that is satisfactory. Fricker seems to have this view, saying "the practice of blame is significantly disunified, and is therefore likely to have distinctive or otherwise central features that may not be present in all instances" (2016: 166). Perhaps, then, we ought to be ecumenical about blame.

I am therefore happy to fall back on a more *moderate* thesis. This, less ambitiously, says that the *Appraisal Theory* presents merely a qualifying standard for blame. Its *Appraisal Requirement* represents a necessary condition, not necessarily a sufficient one. If we instead opt for this moderate thesis, then the objections about the theory's over-inclusiveness and under-explanation are defused (or at least deferred) in a different way. The *Appraisal Theory* now fails to explain blame's familiar properties, and accommodates responses we would not intuitively deem to be blame, because it is an *incomplete* theory. In which case, these objections may well reflect an inadequacy of the theory – it is too thin to amount to a complete account of blame. A more complete theory would add further conditions, specifying additional constitutive or functional properties in order to filter out the counter-intuitive results. (Although again, thickening the theory in this way means it may struggle to account for the full array of blame, with all its attendant variables. We run the risk of the sort of fancy dancing which Shoemaker and Vargas highlight.)

Whether we opt for the moderate or the ambitious thesis, I believe that the *Appraisal Requirement* describes an important and consistent feature of blame, one which I think has sometimes been overlooked.[97] It registers how blame is the response we have when, and because, we appraise people as morally underperforming whilst exerting control over that underperformance. It may well be more than this but, before we get to the question of constitutive or functional properties, we can at *least* say this about blame. Responses which have this feature clearly play a very significant role in our lives – appraisals of this kind are understood to give normative grounds to treat each other in very distinctive and consequential ways. The ambitious version of the *Appraisal Theory*, which I endorse, proposes that *blame* is the natural concept to use to distinguish these responses: it says we should just draw the conceptual boundary here.

At the very least, my conception of blame (and by extension praise) removes any vagueness about how I am deploying these contested concepts. My subsequent use of blame and praise should therefore be understood in these terms (as responses to appraisals of agential features deployed respectively in negative and positive moral performance). This may be a very thin conception of blame and praise. But as I hope

[97] Shoemaker and Vargas, for instance, note how "[n]early all the leading theories of blame agree about at least this much: blame is a response to a person in light of his or her perceived norm violation, where the blamer takes that violated norm seriously" (2019: 582, original italics). I agree. This is tantamount to an appraisal of my first essential constituent of blameworthiness, that of moral underperformance. But I think this misses the second essential constituent. I think nearly all the leading theories of blame would also agree that the blame they describe is a response to a person exercising *control* over their norm violation. (The Fricker-style *blame as judgment of moral fault* is perhaps the notable exception).

will be borne out in Part Two, it is a conception which makes for greater clarity in the analysis of skepticism about basic desert moral responsibility.

2.4. Formulising Basic Desert

2.4.1. A Formula for Theories of Moral Responsibility

I have now finished setting out my overarching conception of moral responsibility. This conception has deconstructed moral responsibility into three essential composite elements. In briefest summary these are agential features; blame (and praise) responses; and the normative connection between these first two things such that the former, when deployed in non-neutral moral performance, contribute to reasons of blameworthiness (and praiseworthiness) in favour of the latter.

This completed conception generates a four-placed formula for how theories of moral responsibility are constructed. I propose we can say that the following is the core claim made by any moral responsibility theory (call this formula MRT):

Due to agential features A deployed in non-neutral moral performance, plus facts B, we have a normative reason of kind C in favour of response D.

Component A represents the two essential constituents of blameworthiness (or praiseworthiness, depending on whether it is a moral overperformance or underperformance). Component C is a reason of blameworthiness or praiseworthiness.[98] Response D is necessarily a blame or praise response, given Component A and the *Appraisal Theory* of blame.

According to the arguments I have offered, Components A, C and D are minimum and essential requirements for any theory of moral responsibility. Component B on the other hand is optional. It is an option taken by those theories which deem deployed agential features to be insufficient to generate a reason of blameworthiness, and thus where additional facts must be specified if deployed agential features are to gain this normative significance. As I noted in Section 1.4.7, however, the vast majority of theories do not appear to take this option. They refer only to the relevant deployed

[98] Note that a *normative reason of kind C* is intended to describe the way in which the balance of reasons is affected, in any direction. This might include the weakening and elimination of existing reasons as well as the addition of new reasons. This becomes relevant in my discussion of basic desert in Chapters 3 and 4.

agential feature in explaining the generation of the normative reason. In which case, they effectively specify component B as nil.[99] Each of the four components A – D can be specified in various different ways. Indeed we have seen that they *have* been specified in various different ways. This, I suggest, accounts for the wide variety of theories of moral responsibility which we find.

I will now finish this chapter by setting up the problem for Part Two of the thesis. My focus now narrows as I turn to consider a distinct genus of moral responsibility: *basic desert moral* responsibility.[100] This kind of moral responsibility was initially distinguished and articulated by those *skeptical* about its existence.[101] For these skeptics, basic desert is a mythical species: no one is responsible in this particular sense.[102]

A chief aim of Part Two is to complete the MRT formula for *the moral responsibility theory which is targeted by basic desert skepticism* (call this SMRT). To be precise, I mean the kind of moral responsibility which the skeptics have sought to articulate in terms of basic desert, and for which they have provided skeptical arguments which purportedly weigh against it. As I will now show, how we should specify SMRT is less clear than we might hope.

2.4.2. Applying the Formula to Basic Desert

The *Standard Characterisation* of Basic Desert

So, what is basic desert? Joel Feinberg is sometimes associated with the development of the idea.[103] In his landmark essay *Justice and Personal Desert* Feinberg contrasts "basic" with "derived" deserts (1970: 82). This is a distinction between two different kinds of thing that can be deserved. For Feinberg, it is only attitudes, such as approval and disapproval, which are basically deserved. Other "modes of treatment", such as

---

[99] The obvious exception here, as discussed earlier, are contractualist theories of moral responsibility – these specify component B in terms of normative facts about what contractualist moral logic requires.

[100] I refer to it as a *genus* in the sense that particular species (theories) of moral responsibility may belong to it, depending on whether they possess its distinctive characteristics. (*Order* or *family* might be a more appropriate label still, given the number of species which feasibly qualify as theories of basic desert!)

[101] For instance, Pereboom and Caruso did important early work in articulating basic desert whilst being avowed skeptics about it.

[102] For concision I will hereafter use the term 'basic desert' to refer (unless stated otherwise) to basic desert *moral responsibility*.

[103] For this see, for instance, Caruso (2016: 26).

punishment, can also be deserved but only insofar as they are expressive of those more fundamental attitudes. So *basic* in Feinberg's sense denotes a primacy, as opposed to derivativeness, amongst those things which can be deserved.

I understand Feinberg's notion of basicness, however, to be importantly different to that which is implicated in basic desert. My impression is that Derk Pereboom was the first to explicitly articulate this idea (I am not aware of Feinberg using the term *basic desert* in this second sense). Here is Pereboom's classic expression of basic desert moral responsibility:

> For an agent to be morally responsible for an action in this [basic desert] sense is for it to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations (2014: 2).[104]

This characterisation has become firmly established in the literature as the *locus classicus* of basic desert. It is frequently echoed by others and endorsed wholesale by fellow skeptics (most notably Gregg Caruso). So, let's call it the *Standard Characterisation*.[105] For instance, Leo Menges offers a characterisation of basic desert which directly borrows from the *Standard Characterisation*. It reads:

> The sense of desert at issue does not depend on positive effects of blaming, such as deterring people from future wrongdoings, or on whether the agents have consented to being blamed in advance. The relevant sense of desert is such that

[104] Pereboom here talks of basic desert for actions. I have already discussed how the scope of moral responsibility is often taken to extend beyond actions to include the consequences of actions, omissions and also attitudes (see Section 1.4.2). There is no obvious reason, at this stage in the discussion, to assume that the formulation which Pereboom sets out for basic desert for actions could not be extended to include these other things.

[105] Other noted moral responsibility skeptics – such as Galen Strawson (1986, 1994), Neil Levy (2011) and Bruce Waller (1990, 2011) – have not expressed their views in terms of basic desert. In my view their arguments might be translated into such terms – at root they are skeptics about essentially the same thing – but I will not seek to defend this claim. Instead, I will focus predominantly on those skeptics who have explicitly presented arguments for skepticism about basic desert. This includes Pereboom and Caruso (occasionally alongside Stephen Morris); my use of 'skeptics' from now on refers specifically to this exclusive club.

> whether agents deserve blame for their actions only depends on the negative moral status of what they did and on how they were when they did it, especially the kind of agency and knowledge they had (2022: 2).[106]

I take Menges's formulation to be essentially equivalent to Pereboom's. It is worth next briefly unpacking what these characterisations tell us about basic desert.[107] This will serve to establish two things. First, what it is that they clearly establish as essential to the concept. But also, where there are gaps in this specification and there is clarificatory work to do to.

Sheer Agential Features

The *Standard Characterisation* and Menges' characterisation allow us, I think, to specify components A and B in SMRT clearly enough. Both formulations describe a reason in favour of blaming or praising which exists strictly owing to facts about the agent. So, *just because she has performed the action* and *how they were when they did it, especially the kind of agency* they had. Additionally, the agent must have performed a *morally wrong* or *morally exemplary* act. Together, then, these are facts I suggest we can readily parse in terms of the agent's *(control-conferring) agential features deployed in non-neutral moral performance*[108]

Furthermore, both formulations describe a condition of exclusivity with regards to the relevant agential features. They say that blame is owed *just*, or *only*, because the agent has exhibited those features. Notably, both Pereboom and Menges explicitly exclude consequentialist and contractualist considerations as bearing upon the relevant reason. But these, I think, are both intended as examples of a more general exclusion

[106] Menges' formulation of basic desert concerns moral responsibility in the case of negative moral performance, so it talks only of blame (and not praise) as that which is basically deserved. This reflects his broader concern to identify precisely what kind of blame, and blameworthiness, skeptics about basic desert are skeptical about. But there is no reason to think his formulation could not be extended to include praise (in the way that Pereboom's formulation explicitly is).

[107] Scanlon also offers a very similar characterisation of (basic) desert, one which he too explicitly introduces as inspired by Pereboom's formulation: "What is distinctive about a desert-based justification for treating a person in a certain way is that it claims that the treatment is made appropriate simply by certain facts about that person or what he or she has done. The 'simply' in this formulation is intended to exclude justifications that appeal to the effects of treating the person in that way and justifications that appeal to the fact that this treatment is called for by some institution (that is itself justified on some basis other than an appeal to desert)" (2013: 101).

[108] Epistemic agential features are also clearly stipulated: the agent must possess *understanding of the moral status* of their action; facts about their *knowledge* are relevant.

condition.[109] What this says is that, in basic desert, we need not be beholden to any particular normative theory in order to appreciate the reason-giving force of agential features. The facts speak for themselves – we need not consider them in light of any wider moral theory or principle.

So, in basic desert, deployed agential features are the sole contributor in reasons of blameworthiness and praiseworthiness. This tallies, I think with the description of the responses in question as *deserved*. Pereboom and Caruso do not provide any independent account or explanation of what they mean by desert, so it is not entirely clear what we should read into this.[110] Whilst I cannot offer an in-depth discussion of desert, the standard view of desert is that a person's desert is based strictly on past and present facts about that person.[111] So, it is reasonable to consider this *agent-centeredness* to be a core commitment of all desert claims. The tightness of this agent-centredness condition might be up for debate. But it seems sure that restricting the reason-generating facts to agential features, as the *Standard Characterisation* of basic desert does, would meet any such condition.

Basic desert moral responsibility, then, implicates what I call *sheer agential features.* By which I mean, it is an agent's deployed agential features alone – *sheerly* – which explains the agent's blameworthiness or praiseworthiness. The agent's deployed agential features therefore have what we might call a *freestanding* normative significance. They are not reliant on additional facts in order to generate a reason of basic desert. This already highlights, I think, the special significance of basic desert. It describes a kind of moral responsibility and blameworthiness which we can (purportedly) acquire simply in virtue of our deployed agential features. Again, as noted before, the majority of theories of moral responsibility appear to claim that the agential feature which they privilege can *alone* (when deployed) account for moral responsibility and blameworthiness. In which case these theories satisfy the *sheer*

[109] This is what is suggested in Pereboom's statement that he is excluding contractualist and consequentialist considerations *for example*. The strong implication being that there are other possible kinds of consideration which would also be excluded. Nonetheless, Pereboom himself consistently namechecks consequentialist and contractualist considerations as those which are excluded. Why is this? My sense is because these are considerations on which two well-established traditions of moral responsibility theory have been built, both of which it is thus worth explicitly distinguishing from basic desert. It is *not* because these are the only normative considerations we need worry about excluding.
[110] Nor does Menges offer a conception of desert – but he is only seeking to echo the *Standard Characterisation*, not (yet) offer his own theory of what basic desert is.
[111] See for example Rachels (1997: 176); Feinberg (1970: 72); Miller (1976: 93).

agential features condition of basic desert (and thus might be in scope of basic desert skepticism).

We can therefore specify Component A and Component B in SMRT, taken together, as sheer agential features deployed in non-neutral moral performance. Where Component B corresponds to no (or nil) additional facts.[112]

The Imprecision of the *Standard Characterisation*

We have clarified components A and B in SMRT. What about components C and D – what is the kind of blameworthiness and blame implicated in basic desert? Here the situation is much less clear. Both the *Standard Characterisation* and Menges' characterisation specify the relevant response as – quite simply – *blame and praise*. Furthermore, neither author (nor Caruso) gives any further specification of what they specifically understand blame and praise responses to be. This, then, is highly unsatisfactory. Given the many and diverse ways that blame (and praise) have been interpreted, as set out earlier in this chapter, it is simply unclear how we are supposed to understand the blame and praise being referred to here.

Again, can we read anything further into the fact that the blame and praise in question is at least specified as *deserved*? Another core commitment of all desert claims, it seems, is that they entail liability to broadly *favourable or disfavourable responses*. For instance, Kleinig writes how "[a]nything which is pleasant or unpleasant can be said to be deserved (by people, that is)" (1921: 72). Feinberg similarly says "[t]he deserved object must be something generally regarded with favor or disfavor even if, in some particular case, it is regarded with indifference by a person said to deserve it" (1970: 61).[113] Again, the basic desert skeptics do not provide any account of how *they* are understanding desert, so we must be careful in assuming too much. But

[112] Very briefly, I wish to register how McKenna (2024) has offered a defence of the harms he associates with blame which he claims amount to a basic desert-based defence. (I mentioned earlier in the chapter how McKenna identifies these harms as the negative effect on friendships; interference from others in one's personal life; and the undermining of emotional stability). McKenna claims these harms are (sometimes) *non*-instrumentally good. Inflicting them can be a way to show that both the blamer and the target care about the moral norms in question. As I discuss in Section 3.3.3, I agree that reasons of non-instrumental goodness are in principle amenable to basic desert. They clearly need not depend on instrumental normative considerations. However, my worry about McKenna's account is that the particular goods he describes are in fact not non-instrumental, and thus cannot be squared with the sheer agential features condition in basic desert. For reasons of space I cannot pursue this line of argument here, but I aim to in separate work.

[113] Also it is fair to say that reasons of desert are generally understood, like reasons of blameworthiness and praiseworthiness more generally, to be *pro tanto* reasons.

provisionally, at least, it seems we can therefore assume that the kind of blame and praise they consider to be at stake in basic desert is that which in some sense (respectively) disfavourably and favourably affects it target.

Beyond this, however, the *Standard Characterisation* leaves much unanswered in terms of how it understands the relevant blame and blameworthiness. What kind of favourable and disfavourable responses do we mean precisely – how are we to understand what *praise* and *blame* denote? Even granting that the reason is one of desert, what are the further properties of this reason? Desert is a famously obscure notion. Basic desert is there sorely in need of clarification in these areas. I am not the first to notice this. Haas comments that basic desert is "under-described". He adds that, beyond his *Standard Characterisation*, "Pereboom has little to say about what basic desert is supposed to be"; it is a notion that is "simply under-developed and in need of clarification" (2013: 229). McKenna likewise remarks that there is "little clarity about what this basic notion of desert comes to" (2012: 117)[114]

Here then is an example of the sort of conceptual imprecision concerning blame (and blameworthiness) which motivates my broader clarificatory project. This imprecision means there is room for *interpretation* in how we should understand the kind of moral responsibility which basic desert skeptics intend to undermine with their skeptical arguments (or: in how we should we understand SMRT).[115] Whilst there is agreement and clarity about the sheer agential features condition, it remains unclear how exactly we ought to specify components C and D in SMRT. It is this problem that I will address in Part Two of the thesis.

[114] For others who highlight the need to clarify basic desert – in particular the senses of blame and blameworthiness which it implicates – see Elzein (2013); Menges (2022); McKenna (2024).

[115] I should note that Menges explicitly acknowledges this imprecision. Indeed this forms the motivation for his own project which aims to more precisely identify the sort of blaming responses the skeptic should be skeptical about (Menges 2022). My analysis of basic desert will draw out points of comparison against his. Also, Pereboom has more recently refined his characterisation of basic desert, as I will shortly discuss in Chapter 3. I choose to begin my discussion with his earlier formulation of basic desert because it is instructive to see how Pereboom has developed the characterisation. Also, the revisions in the new characterisation principally affect the specification of the deserts – blame and praise – that are owed in basic desert. The specification of agential features remains essentially unchanged (aside from a slight modification to the epistemic requirement, which I mention in Section 3.2).

2.5. Concluding Remarks

In this chapter, I have put forward a theory of blame – the *Appraisal Theory* – which completes my overarching conception of moral responsibility. This theory is very thin. It says that any and all blame necessarily involves an appraisal of the essential constituents of blameworthiness (the deployment of agential features in moral underperformance). The theory therefore naturally complements my preceding account of agential features and blameworthiness.

I endorse an ambitious version of the *Appraisal Theory*, which says that the *Appraisal Requirement* not merely sets a qualifying requirement for blame, but it represents a sufficient condition. This theory is able to account for much of what we typically deem to be blame, and to accommodate most extant theories of blame. That said, the thinness of the account brings limitations. It does not allow us to explain some of blame's typical characteristics. If one sees this a serious deficit of the theory, we can instead adopt the moderate thesis, which says that the *Appraisal Requirement* sets only a qualifying requirement.

I finished the chapter by turning to consider a specific kind of moral responsibility - basic desert. I outlined what we can confidently say about it: we can specify the deployed agential features requirement in terms of sheer agential features. But the *Standard Characterisation* is imprecise when it comes to the blame and blameworthiness which is at stake. This therefore provides a bridge to Part Two of the thesis, which takes up the challenge to discover more precisely the kind of blame and blameworthiness implicated in basic desert.

My arguments in this first part of the thesis may seem a bit like treading over somewhat familiar ground – many of the concepts I have discussed in these first two chapters are very well-entrenched in the moral responsibility debate. But the fact they are entrenched does not mean they are easy to interpret – in fact the opposite is often true. I have therefore sought to martial these conceptual ingredients in what I hope is a new and clarifying way.

PART TWO: REASSESSING BASIC DESERT SKEPTICISM

Chapter 3 - The Moral Responsibility at Stake in Basic Desert Skepticism: Three Candidates

3.1. Introduction

Recall my four-placed formula for the core claim made by theories of moral responsibility (MRT): *due to agential features A deployed in non-neutral moral performance, plus facts B, we have a normative reason of kind C in favour of response D*. The central aim of Part Two of this thesis is to complete this formula for a distinctive kind of responsibility: that which basic desert skeptics have given us cause to doubt (SMRT). At the end of Chapter 2 I provided this specification for components A and B, the kind of facts on which this kind of moral responsibility is premised. These components taken together are understood in terms of agential features deployed in non-neutral moral performance, without any additional (component B) facts. Or what I refer to as *sheer agential features*.

I now turn to components C and D as these figure in SMRT. These components together specify a kind of normative significance which said sheer agential features might possess: how it is they generate a reason of blameworthiness or praiseworthiness, and thereby alter the balance of reasons with regards to a specified category of blaming or praising response.[116] How should the basic desert skeptic specify these two components in SMRT? As I will show, this has remained obscure.

I am to clarify things, in the first instance by using this chapter to set out three views. Each view builds upon the stipulation of sheer agential features, but each offers a different specification of the two components in question (C and D). In my discussion of these three views, I will again predominantly (although not exclusively) focus on

[116]To reiterate, 'altering the balance of reasons' is meant inclusively, capturing the weakening or elimination of existing reasons as well as the addition of new reasons. Also, I here use blame and praise, and blameworthiness and praiseworthiness, according to the conceptions I developed in the first two chapters. So, for instance, blame is any response had to an appraisal of agential features deployed in negative moral performance. Blameworthiness provides a normative reason to respond with blame, which is at minimum grounded in agential features deployed in subpar moral performance. It is worth stressing that my analysis of the three views I set out in this chapter does not require accepting these conceptualisations, ones which I recognise might be controversial. Nonetheless I do aim to impress how conceptualising blame in such thin terms – so thin that it effectively renders blame and praise-talk disposable – is preferable, at least for the purposes of understanding basic desert skepticism and the specific kind of responses it impugns.

what they say about blame and blameworthiness.[117] In doing so, I will refer to Component C as the 'Blameworthiness-Component' and Component D as the 'Blame-Component' (but note that we might instead use Praiseworthiness-Component and Praise-Component, where we are instead concerned with specifying SMRT with regards to praise). What most clearly distinguishes each of these views is their different interpretation of Blame-Component: the nature of the blame response which they take skepticism to undermine. In this regard, these views sit on a spectrum – from narrower to wider inclusivity – in terms of the relevant kind of blame and praise.

The first two views I will discuss are existing, competing conceptions of basic desert. First, discussed in Section 3.2, is basic desert as the skeptics themselves – most notably Derk Pereboom and Gregg Caruso – understand it. I aim to render this view more clearly than it has so far been stated. These skeptics provide a distinctly retributivist account of SMRT, and it is an account in which (at least for Pereboom) the idea of non-instrumentally good harm and benefit is integral. But it is a view with residual vagueness concerning its use of the concepts of blame and praise. Section 3.3 involves a sustained discussion of the second view: a refined interpretation of basic desert developed by Leo Menges. This view understands basic desert in terms of a reason to respond in a way which would otherwise violate a moral claim held by its target, but which does not in fact do so because the relevant claim has been forfeited. I summarise Menges' view, and also the desiderata which he uses to defend it. I then critique Menges' arguments, first by defending the skeptics' interpretation of basic desert against his criticisms. When it comes to the interpretation of Blameworthiness-Component, I argue that the skeptics' view in fact provides a more promising option for a basic desert theorist than Menges' own proposal, and even on the terms of his preferred criteria. Then I argue against Menges' central claim: that his account describes that specific kind of moral responsibility which basic desert skeptics' ought to be concerned with. I argue that the desiderata he uses are inadequate for adjudicating this. It thus remains an open question how we should conceive of basic desert.

[117] This narrowing of focus is again foremostly for reasons of space. But it is also because all three of the views canvassed in this chapter have clear application to blameworthiness, and so can be compared on this basis. Whereas it less obvious how the second view I discuss – Menges' account – applies to praiseworthiness.

Finally, in Section 3.4, I sketch a third view, one which is maximally inclusive with regards to the relevant blaming response (component D in SMRT). Unlike the other two views, this should not be understood as a theory of basic desert. Instead, I call this *Basic Reason* moral responsibility. I will highlight some theoretical and practical benefits we would enjoy were it to turn out that the skeptical considerations which have been mounted against basic desert specifically discredit this version of moral responsibility, rather than the other two canvassed. But again, this remains an open question, one to be settled in Chapter 4.

3.2. Basic Desert in the Hands of Skeptics – *SMRT: Retributivism*

My aim is to establish the precise nature of that species of moral responsibility which basic desert skepticism brings into question. The natural place to start is with the skeptics' own description of it. In particular, we should start with the account given by Derk Pereboom, whose *Standard Characterisation* of basic desert has become established as the definitive statement of the concept. Yet in the previous chapter we registered, as have others, the imprecision of this characterisation, with regards to both the relevant kind of blame and the relevant kind of blameworthiness (the reason for such blame). This imprecision means it remains unclear how basic desert skeptics, particularly Pereboom, would fill out SMRT. Nonetheless I aim to show we can read Pereboom as more definitive – and more aligned with his fellow basic desert skeptics such as Caruso – in his specification of SMRT than some have taken him to be.

3.2.1. What *SMRT: Retributivism* says about Blame

Take blame first. We have seen how Pereboom's *Standard Characterisation* describes the relevant response-type as – quite simply – *deserved blame and praise*. And, having established that liability to *(dis)favourable responses* are an essential property of any desert claim, we arrived in Chapter 2 at a provisional, at least, picture of the kind of blame and praise at stake in basic desert. Namely, that *which we can expect to favourably or unfavourably affect its recipient*.

This picture is reinforced by a new, slightly updated version of the *Standard Characterisation* which Pereboom gives in his most recent book. He now says that:

> For an agent to be morally responsible for an action in the basic desert sense is for the action to be attributable to her in such a way that if she was sensitive to

> its being morally wrong, she would deserve to be blamed or punished in a way that she would experience as painful or harmful, and if she was sensitive to its being morally exemplary, she would deserve to be praised or rewarded in a way that she would experience as pleasurable or beneficial. The desert at issue is basic in the sense that the agent, to be morally responsible, would deserve such blame or punishment, praise or reward, just by virtue of having performed the action with sensitivity to its moral status, and not, for example, by virtue of consequentialist or contractualist considerations (2021b: 11).

The apparently significant development here is the stipulation that the relevant blame must be *experienced* in a certain way: *as either painful or harmful*, and praise as *pleasurable or beneficial*.[118] This qualification seems to add a new degree of specification to the blame and praise in question. But, if we accept that liability to (dis)favourable responses is an essential property of any desert claim, then Pereboom's updated characterisation reinforces but in fact adds relatively little to the original. A response which causes one to experience pain or harm, or pleasure or benefit, is a response which entails that one has been favourably or unfavourably affected.[119] So, Pereboom is effectively here making explicit the property of (dis)favourableness which we had already taken to be implicit.

Pereboom's updated characterisation however does now specify *experienced pleasures* and *pains* as belonging to the relevant response-type. This seems to pinpoint a very particular kind of (dis)favourableness as being what is at stake. Whereas the category of (dis)favourable responses – those that can be good or bad for people, the stuff of desert – plausibly ranges more widely than merely those experienced as pleasurable or painful. At least, it does for anyone who does not subscribe to a wholly hedonistic theory of wellbeing. For instance, a desire-satisfaction theorist might regard a response which fulfils an individual's desires as giving that individual a distinctive

[118] The other innovation in this new version is the weaker epistemic condition, where mere sensitivity to (rather than understanding of) the action's moral status is required. As discussed, this condition is of less relevance to my arguments.

[119] Strictly speaking, and in line with the discussion of desert in the previous chapter, basically deserved blame and praise are responses which we expect to or typically have harmful and beneficial effects, rather than ones which necessarily have such effects in any given case. I think we can understand Pereboom here in these terms (so, the basically deserving agent would warrant the actually harmful blame they experience in the world in which there are no unexpected or atypical impediments to its being experienced as such).

kind of good. A case which uses the "experience machine" famously described by Robert Nozick (1974) illustrates this. For instance, suppose both Bjorn and Jimmy have a strong desire to go to Wimbledon and watch the tennis final at Centre Court. Roger wishes to reward both Bjorn and Jimmy for their recent good behaviour. In Bjorn's case, he provides him with a ticket for the final. There are no more tickets available, however, for Jimmy. So, Roger instead borrows Nozick's experience machine and, plugging Jimmy in, allows him to experience going to Wimbledon and watching a final in exactly the same way as Bjorn experiences it. Both agents therefore have the same equally pleasurable experience, yet only Bjorn's desire to actually attend the Wimbledon final is satisfied. A conception of basic desert which understands the relevant deserts solely in terms of experienced pleasure and pain could not account for the satisfaction of Bjorn's desire as a response which he can basically deserve.

So, Pereboom's new 'experienced pain/pleasure' stipulation is curious and worth noting. But, whilst I will occasionally return to it in what follows, I will not dwell on it. First, because it not a claim which he anywhere further elaborates or defends. Second, because his updated *Standard Characterisation also* stipulates responses experienced as *harmful* and *beneficial* as the relevant response-type, along with pain and pleasure, and this seems to hedge things. Harms and benefits, unless specified, usually refer more generally to a wider category of (dis)favourable responses – inclusive, for instance, of the sort which concerns the desire-satisfaction theorist as well as the hedonist. So, it remains reasonable to interpret Pereboom as holding that basic desert concerns responses which are (dis)favourable in the more general, not-necessarily-hedonistic sense.

Interestingly, Pereboom describes the updated *Standard Characterisation* as "more decisively retributivist" than its predecessor (2021b: 12, Fn.12). However, on its face, this seems questionable. A full analysis of retributivism is well beyond my scope. But at minimum retributivism involves the idea that punishment – which necessarily involves the intentional infliction of harm – is sometimes owed solely for the reason of desert.[120] As leading retributivist Michael Moore puts it, for instance: "Retributivism is the view that we ought to punish offenders because, and only

[120] See Bedau and Kelly (2015) for an overview of punishment.

because, they deserve to be punished. Punishment is justified, for a retributivist, solely by the fact that those receiving it deserve it" (2010: 153)[121]

Pereboom's updated *Standard Characterisation* explicitly provides that the blame and praise responses implicated in basic desert are – in virtue of being deserved responses – responses which inflict harm (disfavour) and benefit (favour). But it does not stipulate that those responses must be *intended* to produce their harm or benefit. Moreover, whilst deserved responses may often involve an intention to harm or benefit, it is not necessary that they do.[122] For instance, perhaps one's expression of sadness provokes pain in another. Even if the sadness response includes no intention to inflict harm, it is conceptually coherent that the provoked pain might yet be deserved. More widely, many experiences we might naturally deem to be deserved are not the product of any interpersonal response at all (an athlete guilty of doping has a stumble which robs them of a gold medal; the honest athlete behind them is fortunately promoted into first place). So, given that deservedness alone does not necessitate liability to intentional harms and benefit, the updated *Standard Characterisation* does not decisively deliver a retributivist picture. It omits the property of intentional harmfulness which is the hallmark of retributivism (and likewise intentional beneficence).[123]

Admittedly, punishment or reward are responses which *do* seem necessarily to entail the intention to harm and benefit respectively. But it is Pereboom's inclusion of *blame* as well as punishment, and *praise* as well as reward, in the *Standard Characterisation* which continues to muddy things. If these are meant to signify more than punishment and reward, then (again) it remains unclear what they signify. If not, then they seem superfluous – why not dispense with them altogether, and express the relevant responses in basic desert simply in terms of punishment and reward? So, uncertainty

[121] Concordant accounts of retributivism can be found in Kant (1797), Berman (2008), Kershnar (2000), and Husak (2000).

[122] As we will see in Section 3.3, Menges exploits this very conceptual possibility – deserved blame without intention to harm – in developing his alternative account of basic desert.

[123] Another sense in which the label 'retributivist' is not entirely apt for the updated *Standard Characterisation* is that retributivism concerns the justification of punishment and blame only, rather than praise and reward. But whilst there is no analogous terminology readily available in the praising case, the relevant ideas – *intentional* benefit (rather than harm) *on the basis of desert* – can be readily translated across.

remains about how we are to interpret the relevant blame and praise in SMRT, as the skeptics see it.

Nevertheless, if we consider what basic desert skeptics have said elsewhere, we can give some more definition to the kind of blame and praise they have in mind. Most notably, Pereboom and Caruso consistently describe themselves as skeptics only about moral responsibility of a distinctive sort: *accountability*. As Pereboom says, "it's the accountability sense, conceived specifically as invoking basically deserved pain or harm, that separates the parties in the debate" (2021b: 14). Accountability here is understood as the practice of holding each other (and ourselves) to moral demands through sanctions and rewards, which is precisely to say the intentional imposition of harms and benefits. For instance, Pereboom and Caruso approvingly cite Shoemaker's definition of holding accountable. In the case of wrongdoing, it means "precisely to sanction that person, whether it be via the expression of a reactive attitude, public shaming, or something more psychologically or physically damaging" (Shoemaker 2011: 623).[124] This view is a descendant of Gary Watson's classic sanctions-based theory of accountability (Watson 1996). So, it is accountability which is taken to be distinctively undermined by disbelief in basic desert.[125] Other senses of responsibility – answerability and attributability, for example – are seen as untouched by basic desert skepticism.[126] This gives further definition to how these skeptics are understanding blame and praise: whatever other characteristics these responses have, they are identified as accountability responses of sanction and reward. This supplements the updated *Standard Characterisation* with the missing condition of intentionally (dis)favourable responses so that we arrive, decisively, at a retributivist conception of basic desert.

This retributivist interpretation of Pereboom brings him into line with the other prominent basic desert skeptics who have put things more explicitly and exclusively

[124] In more recent work Shoemaker himself has moved away from this sanctions-based understanding of accountability. He instead argues that *confrontation* is accountability's defining feature – and unlike sanctioning, confrontation does not necessarily involve harming or benefitting its target (see Shoemaker 2015: 107). But what is important for my purposes is that the basic desert skeptics have consistently understood accountability in terms of sanction (and reward).

[125] Specifically, accountability *on the basis of sheer agential features*. Accountability practices can be, and have been, instead justified on other grounds (such as contractualist or consequentialist considerations). These brands of accountability are clearly not in scope of basic desert skepticism.

[126] See Caruso and Pereboom (2022: 5-7) for discussion which confirms this.

in retributivist terms. Most obviously Caruso and Morris (2017), who expressly identify basic desert with 'Retributivist Desert Moral Responsibility'. Contrasting their view with Strawsonian theories of moral responsibility which focus exclusively on the appropriateness of reactive emotions, they say "it seems appropriate in the context of the free will debate to reject the understanding of moral responsibility favored by P. F. Strawson in favor of one in which the propriety of reward and punishment plays a central role" (Caruso and Morris 2017: 843–44).

Some moral responsibility skeptics who do not couch their skepticism in terms of basic desert have similarly stated it is retributive moral responsibility which they specifically oppose. Galen Strawson, for example, summarises his view as follows: "We are what we are, and we cannot be thought to have made ourselves *in such a way* that we can be held to be free in our actions *in such a way* that we can be held to be morally responsible for our actions *in such a way* that any punishment or reward for our actions is ultimately just or fair" (1994: 15, original italics). Indeed, the kind of free will he rejects is one which makes intelligible deserved eternal suffering in hell, a particularly extreme form of punishment (Strawson 1994: 9).

So, we can now reasonably place Pereboom in this anti-retributivist camp and consider the basic desert skeptics – he, Caruso and Morris – to hold a unified view of the response-type implicated in basic desert. This view says that *intentional harms and benefits* are the relevant responses: this is, at least partly, how they specify the Blame-Component in SMRT.[127] But the question remains: why not characterise the responses

[127] This interpretation of Blame-Component gives rise to an apparent tension. The propriety of reactive attitudes (such as anger and resentment) is undoubtedly a core part of what basic desert skeptics are concerned to debunk. Caruso and Morris, for instance, state that free will skepticism concerns "the justification of various sorts of desert-based judgments, attitudes, or treatments—e.g., those related to just deserts, praise and blame, *and the reactive attitudes… of resentment, indignation, guilt, and righteous anger*" (2017: 840, italics added). Pereboom has consistently waged against the appropriateness of the reactive attitudes, devoting his latest book *Wrongdoing and the Moral Emotions* to the issue (2021). However, it seems generally agreed that the reactive emotions are not issued with any intention to harm – emotions are typically understood to be *involuntary* (e.g., Müller 2021). How then do we square the skeptics' opposition to these attitudes with (what I interpret as) their exclusive interest in intentionally harmful responses? For Pereboom and Caruso, the answer has traditionally been that these emotions necessarily involve a *presupposition* that their recipient deserves to be on the receiving end of the expression of the emotion (Pereboom 2021: 2). So, any episode of anger necessarily involves the presupposition that the offender *would* basically deserve to be the recipient of expressed anger – more precisely, a recipient of its harmful content. This is a presupposition which, as basic desert skeptics, they think is necessarily mistaken. So, as I read the skeptics here, the claim is not that the reactive attitudes are themselves the kind of thing which might be basically deserved. Rather, it is subtly different: they are necessarily unfitting or irrational because they mistakenly attribute to their target a property which does not exist, namely that of basic desert – the appropriateness of intentional harm.

simply in these terms, rather than further qualifying them as responses of blame and praise? Is it punishment and reward of *any* flavour which is the currency of basic desert, or does the fact they are construed as blame and praise responses delimit them further, picking out *particular* harms and benefits? No clear answer is forthcoming. The unresolved ambiguity here means it is difficult to be absolutely definitive about how these skeptics specify the Blame-Component.

Am I being properly charitable to Pereboom and his fellow skeptics here? Is there really no more refined conception of blame to be found in their work, one which would allow us to definitively specify the responses about which they are skeptical? Perhaps surprisingly, there is not. We do not find in the work of avowed basic desert skeptics such as Pereboom and Caruso any theory of blame (or praise) which gives further definition to the kind of blame (and praise) they aim to debunk.[128] Pereboom's characterisation of basic desert – the original and then updated version – effectively gives the full specification. This is why it warrants such close inspection. And this characterisation, as discussed in Chapter 3, preserves the conceptual vagueness I argue is endemic to blame- and praise-talk. As such it remains unclear to me what explanatory work the concepts of blame and praise are intended to do – how do these further specify the relevant response-type, beyond being deserved harms and benefits? Perhaps they are not intended as explanatory.[129] Pereboom and Caruso might be conceptualising blame in a sense similar to the 'thin' sense I have advocated, where this simply designates that the response is one given to an appraisal of agential features deployed in negative moral performance, without attributing any further content to the

This is why basic desert skepticism undermines these emotions – they are "doxastically irrational" (Pereboom 2014: 128). However, in his latest book Pereboom, at least, is less committed to this claim about the representative content of reactive attitudes (2021: 33). This leads him to develop alternative arguments against moral anger, which instead concern its damaging instrumental effects. (In Chapter 4, Footnote.209 I briefly outline a different kind of misrepresentation which I think is involved with reactive attitudes such as moral anger).

[128]The same I think is broadly true of that wider group of moral responsibility skeptics – notably including Galen Strawson (1986; 1994), Neil Levy (2011) and Bruce Waller (2011) – who, whilst holding views broadly aligned with Pereboom and Caruso, have not articulated their skepticism in terms of basic desert. But I will remain predominantly focused on basic desert skeptics (and, in Chapter 4, on the kind of argument for skepticism which they in particular favour).

[129] Although note that Pereboom occasionally drops the blame and praise qualifiers altogether, such as: "for an agent to basically deserve pain or harm she must have a kind of control in action, a type of free will, that is unavailable to us" (2021b: 31). This would seem to support this 'non-explanatory' interpretation.

nature of the response. But we have no firm grounds to attribute this nor any other precise conception of blame and praise to Pereboom and his fellow skeptics.

Nor do we find in these skeptics' accounts any independent explicit theory of *desert* which might serve to further specify either the kind of blame, or blameworthiness, they take to be at stake. Rather, as I interpret Pereboom and Caruso, it is simply because they understand the relevant phenomenon to exhibit the two core properties of all desert claims I mentioned earlier – liability to (dis)favourable responses on the basis of agent-centred facts – which explains their framing of the phenomenon in terms of desert.

It is this conceptual vagueness, which lies at the heart of existing accounts of basic desert skepticism, that motivates my clarificatory project. This vagueness is most clearly revealed by the way in which the skeptics frame their skeptical arguments. As will become explicitly clear in the following chapter, when I specifically consider Pereboom's manipulation argument, there are no more fine-grained assumptions about blame or desert which skeptics think we need to accept in order to appreciate the force of the argument. (More incidentally, this very same vagueness is recognised by others – including Menges (2023: 616), inspiring his own, similar clarificatory project to which I will shortly turn.)

Nevertheless, despite the residual conceptual vagueness concerning blame and praise, we have at least clarified the following about the position held by prominent basic desert skeptics Pereboom and Caruso. They are skeptical that sheer agential features can provide a reason for *intentionally harmful blame and intentionally beneficial praise*. This is how I understand them as cashing out component D in SMRT.

3.2.2. What *SMRT: Retributivism* says about Blameworthiness

So much for the kind of blame which basic desert skeptics have in their sights. What about blameworthiness: how would they specify the Blameworthiness-Component in SMRT? What properties does this reason to inflict such retributive blame (and its praising analogue) possess, aside from its being grounded in sheer agential features?[130]

[130] Of course, since we are currently interested how basic desert *skeptics* would answer this question, these will be exactly those properties which they insist fail to obtain.

Here I think we can be briefer – because basic desert skeptics have (until recently) had relatively little to say.[131]

That the relevant normative reason is presented by skeptics as one of desert at least tells us *something*, but little that is determinative. The *Standard Characterisation* of basic desert ensures we are working with a strictly non-instrumental notion of desert: the reason is generated purely from deployed agential features, rather than instrumentalist (or any other) considerations. But this leaves it open what further, positive properties this reason possesses. Menges captures this openness well: “do basic desert claims entail that it would be good, reasonable, fitting, justified, permissible, obligatory, or something else to blame or praise, or do desert claims involve primitive evaluations?” (2023: 616). The skeptics’ answer to this question (as Menges likewise recognises) has remained elusive – precisely because their descriptions of basic desert have not been supplemented with any more specified theory of desert.

That the basic desert skeptics I have considered are (now) unanimous in explaining basically deserved responses in retributive terms does not shed much further light. Since retributivism draws upon the conceptual resources of desert, the same openness applies with regards to the properties of the retributive reason. Thus the blameworthiness implicated in basic desert, at least as articulated by these skeptics, has remained amenable to various kinds of specification of the sort Menges lays out.[132]

Yet in his very latest work, in addition to presenting his refreshed characterisation of basic desert, Pereboom (for the first time) describes the relevant blameworthiness in a more specified sense. One which attributes a distinctive axiological property to the retributive reason: they are reasons of *non-instrumental goodness*. He says “we can add that the imposition of such basically deserved pain or harm, pleasure or benefit, is

[131] For all three conceptions of SMRT discussed in this chapter, at least one idea is certainly shared when it comes to blameworthiness: it is understood to constitute only a defeasible *pro tanto* reason in favour of blame. This is in line with how blameworthiness is typically understood by theories of moral responsibility (as discussed in Chapter 1).

[132] The point is not that skeptics are *obliged* to attribute a more specific set of properties to the relevant blameworthiness-reason; perhaps they would be correct to leave things open in this way. As I argue in the next section, this will turn on the force of their skeptical arguments: how specific is the reason-type against which these arguments have force? Ultimately I argue that the skeptics would indeed be well-minded to remain decidedly *non*-specific here, as they have tended to be (at least until Pereboom’s recent claims concerning non-instrumental goodness, discussed next).

non-instrumentally good, in accord with… Michael McKenna's (2012, 2020) specifications" (2021b: 29).[133] Also, in reference to his preferred instrumentalist conception of blame, Pereboom notes that the "rejection of the pain of blaming as non-instrumentally good is significant, and lies at the core of what motivates the forward-looking account" (2021b: 37). So, any ambiguity now seems resolved. Pereboom commits to a particular understanding of the blameworthiness-reason he is skeptical about: it is a reason generated by the supposed non-instrumental value of deserved harm. In doing so, he goes further than his fellow basic desert skeptic Caruso (and Morris). I do not find any such clear commitment in their work – they leave it open what the blameworthiness implicated in basic desert precisely entails.[134]

This affiliates Pereboom's view with a popular interpretation of retributivism. For instance R. Jay Wallace – though he acknowledges it is a narrow and particular interpretation of the concept – describes retributivism as "the view that it is intrinsically good that wrongdoers should suffer harm, and that we therefore have a positive moral obligation to inflict such harms on them" (1994: 60, n. 13). Lawrence Davis, Mitchell Berman and Michael Moore are, for instance, notable retributivists who have understood retributivism in terms of non-instrumental goodness.[135]

[133] Although note that McKenna's account of basic desert is non-retributivist. He is interested in defending particular kinds of harmful blame response, but he is explicit that they need not be *intentionally* harmful (see McKenna 2012: 145-146). To get to retributivism, we must also add in Pereboom and Caruso's exclusive concern with accountability and its commitment to intentional harming and benefitting.

[134] Menges (2023: 626-627) apparently disagrees here. He interprets Caruso and Morris as offering a distinct understanding of the blameworthiness entailed by basic desert – namely as a reason that *justifies* blame. True, were they claiming that reasons of basic desert *necessarily* render (retributive) blame justified, this would indeed be a distinctive view (certainly one different to Pereboom's). But, as far as I can tell, Menges is simply mistaken to attribute this view to them. Indeed, they overtly reject it: "there is no contradiction involved in believing that an agent deserves [in the basic desert sense] to be blamed or punished, but since blaming or punishing her would be undesirable for other more important reasons, we ought not to do so all things considered" (Caruso and Morris 2017: 842). Which is to say, blame would be unjustified despite the agent's basic desert. And, citing Sofia Jeppsson, they stipulate that on their view basic desert entails that "*in the absence of sufficiently strong counter-veiling reasons*, blaming her is justified if her action was wrong" (Jeppsson 2016: 683, original italics; Caruso and Morris 2017: 842). So, this simply amounts to the claim that basic desert provides a defeasible consideration of blameworthiness – one *potentially capable* of justifying blame, but only in fact doing so where it is not outweighed by contrary reasons. This reflects the typical view of reasons of blameworthiness as found in theories of moral responsibility (see Chapter 1), and likewise of reasons of desert (see Chapter 2). So, it is *not* a claim which notably distinguishes Caruso and Morris's conception of basic desert, and certainly not from Pereboom's conception.

[135] Well, Wallace and these others tend to talk of *intrinsic* goodness, a concept often unhelpfully elided with non-instrumental goodness (see McKenna (2024: 65-67) for discussion, drawing on Korsgaard's (1983) classic discussion of this distinction). But we can reasonably take them to mean what is more

In summary, we have ascertained that chief basic desert skeptics Pereboom and Caruso hold the following view. For both, basic desert skepticism targets the existence of retributivist reasons – of blameworthiness and praiseworthiness – to inflict intentionally harmful blame and beneficial praise on the basis of sheer agential features. And, for Pereboom at least, these are reasons which are explained in terms of the non-instrumental goodness of such harms and benefits. This brings some clarity to how they would express components C and D in SMRT. Although not complete clarity, given their conception of basic desert still deploys problematically unclarified concepts of blame and praise. Call this view (including Pereboom's commitment to non-instrumentally good harm and benefit) *SMRT: Retributivism*. I return to assess the merits of this view in Sections 3.3.3 and 3.3.4 below.

3.3. Refining Basic Desert: *Claim Forfeiture*

I now move to consider a different account of basic desert, one recently proposed by Leo Menges. This view merits close attention for several reasons. It is one of very few developed theories of basic desert in the literature, and it explicitly aims to address the same vagueness which motivates me regarding how its skeptics have articulated the concept.[136] Menges is therefore a fellow traveller: he shares precisely my goal of clarifying the kind of moral responsibility implicated in basic desert skepticism. His is also a theory which contrasts with both *SMRT: Retributivism* and the third view I discuss (in Section 3.4), so it is a very useful foil.

Whereas my aims in relation to the skeptics' position were mostly interpretative, in relation to Menges' account they are critical. After briefly setting out Menges' account, I do three things. First, argue that *SMRT: Retributivism* can withstand the objections which Menges directs at it. Second, argue that Menges' own positive proposal for the blameworthiness relevant to basic desert (Component C) is flawed.

aptly described as non-instrumental goodness – given their view is set in opposition to the idea that retributive responses are good with reference to other goods which they (instrumentally) produce. Also, Wallace's claim that retributive reasons endow 'positive obligations' to inflict harm is controversial. Some retributivists claim that desert is merely necessary but not sufficient to justify retribution. This difference defines the debate (one which I need not enter) between 'strong' and 'weak' retributivism, concerning the deontological force of the reason in question. See Berman (2016) for discussion of these alternative views. Again – this is not Pereboom or Caruso's view (see previous footnote).

[136] Though as discussed at the end of Section 3.3.2, Menges takes the skeptics to be vaguer than they in fact are.

Third, and I think most troublingly for Menges, argue that the criteria which he uses to theorise about basic desert are not fit for purpose – we need some other measure if we are to decisively establish SMRT.

3.3.1. Menges' Theory of Basic Desert

Here is a summary of Menges' view.[137] He sees blameworthiness in the basic desert sense as "fragmented": it is a composite of two things. First, the blameworthy agent's forfeiture of a moral claim not to be subject to blaming response X, due to their sheer agential features being deployed in negative moral performance.[138] Second, the existence of some positive reason which favours blaming response X. Menges leaves it open what this positive reason might be. Defending any particular candidate reason – McKenna-style non-instrumentally good harm, say – is a task for the basic desert advocate, and would require a normative argument.[139] Menges' point is conceptual; it is simply that basic desert demands there be *some* such reason (Menges 2023). Together this composite is how he interprets Blameworthiness-Component in SMRT – *this* is how being basically deserving alters the balance of reasons with regards to blame.

Menges (2022) separately develops a complementary conception of the relevant *blame* which is owed in basic desert. This, as we might expect, is blame which would violate a moral claim possessed by its target, were that claim not already forfeited, waived or overridden. And for Menges, what unites such interpersonal claims is that they concern treatment which non-trivially negatively affects us. As he puts it, "we seem to have standing—though in most cases overridable and waivable—claims against others that they not respond to what we do in ways that are nontrivially negative for us" (2022: 9). So, if we all have a standing claim not to be poked in the ribs, then

[137] Note that Menges focuses exclusively on blame and blameworthiness as they figure in basic desert; he does not consider praise and praiseworthiness.

[138] Of course, this is expressed in my preferred terminology of sheer agential features, whereas Menges (as we saw in Section 2.4.2) expresses the same idea in different terms: "because they did something morally objectionable and because of how they were when they did it" (2022: 10).

[139] Menges himself neither endorses nor denies skepticism about basic desert (he has also confirmed his agnosticism here in personal correspondence). His aim is simply to establish the conceptual entailments of basic desert properly understood. (Though he does suggest one kind of positive reason to blame which he thinks is plausible and could be accommodated within his theory of basic desert, a suggestion which I criticise in Section 3.3.3).

blame which takes the form of rib-poking is precisely the sort of blame to which basic desert applies. This is how he interprets component D in SMRT.

For Menges then, basic desert entails a positive reason to blame in a way which would normally violate a standing claim the target has against us. But the action for which there is this reason does not in fact violate any claim, because the reason also entails the forfeiture of that very standing claim. Call this view *SMRT: Claim Forfeiture*.[140]

How does this compare to *SMRT: Retributivism*? In their interpretation of Blame-Component, the two views have some affinity. After all, *SMRT: Retributivism* understands Component D in terms of intentional harms, and intentional harms are very plausibly a kind of response which we each have a standing claim against – they negatively affect their recipient, and typically in a non-trivial way. So, blame in the form of intentional harms is implicated in both accounts.

But as Menges emphasises, the range of responses which non-trivially negatively affect their target, and thus meet the threshold of claim-violation, encompasses more than merely intentional harms. Many other responses feasibly meet this threshold. Menges gives the examples of aggressive reproaches – such reproaches are not typically issued with harmful intent, yet they are nonetheless responses which "intuitively" (in Menges' words) we each have a standing claim against, since they non-trivially negatively affect us. Furthermore, Menges notes how we plausibly have moral claims against responses which apparently do us no harm at all, but which still non-trivially negatively affect us because they infringe our autonomy or integrity. (For instance, being unknowingly touched whilst unconscious). Such responses too, then, are in scope.[141]

---

[140] Neil Levy offers what I take to be a very similar characterisation of the kind of moral responsibility which is at stake, albeit one not couched in terms of basic desert: "Someone who holds that agents can be morally responsible for wrongful acts or omissions … is committed to holding that such agents no longer deserve the (full) protection of a right to which they would otherwise be entitled: a right against having their interests discounted in consequentialist calculations" (Levy 2011: 3). I take this to be a consequentialism-inflected version of the same idea. That idea is that the kind of moral responsibility which is the subject of their skepticism involves being liable to treatment which one is normally morally protected against. (What the Levy quote omits is reference to sheer agential features, and nothing else, as being the grounds for this forfeiture of protection. This is the other essential and generally accepted requirement of basic desert moral responsibility, one which Menges clearly stipulates).

[141] Also, the range of implicated responses would certainly be wider in *SMRT: Claim Forfeiture* than the merely *experienced* harms which Pereboom's updated *Standard Characterisation* (at least partly) invokes. We can plausibly violate moral claims by harming people in ways they do not necessarily

So, *SMRT: Claim Forfeiture* understands Blame-Component more inclusively than *SMRT: Retributivism*. It includes intentional harms but plausibly significantly more besides. In another respect, though, the range of responses implicated in *SMRT: Claim Forfeiture* is more limited, since it is only harms of a certain magnitude which plausibly violate our standing claims (hence Menges' 'non-trivial' requirement). Whereas *SMRT: Retributivism* is concerned with intentional harms, but neither Pereboom nor Caruso place any limit on the *degree* of harm which must be met if it is to be basically deservable.

One thing I wish to register about Menges' account – which he does not emphasise – is what we might call the inherent *variability* in how it specifies Blame-Component. By this, I mean that precisely how we specify the range of responses against which we all have standing claims – those which 'nontrivially negatively affect us' – will surely vary dependent on one's preferred moral theory. For instance, a Kantian will presumably arrive at a different set of standing claims to a utilitarian. Moreover, for at least some moral theories, whether or not a particular response violates an interpersonal claim might well vary further in response to circumstances.[142] For instance, a utilitarian will establish interpersonal claims – at least some of them – with reference to what arrangement of claims best promotes utility. But a change in circumstance might recommend a different arrangement of claims, thus altering the range of responses which can be subject to basic desert. So, this variability means that if Menges' view is correct, then we cannot decisively identify the kind of blame implicated in basic desert until we have settled on our preferred moral theory, and accounted for any circumstantial variability in the moral claims which it establishes. (I will return to this feature of Menges' specification of Blame-Component – its inherent variability – towards the end of Section 3.4).

---

experience (at least directly or knowingly). For instance, we might be harmed by the slander of another, whilst remaining entirely ignorant of it.

[142] This point about contextual variability pertains to McKenna's account too. I mentioned in Chapter 2 that McKenna is confident that the harms involved in interpersonal blame are only mild. He describes the relevant harms in terms of how one's emotional stability, friendships and freedom from interference are undermined (2024: 36). But I am suspicious that we can talk *uniformly* of how harmful such things are, as McKenna tends to. Plausibly the degree of these harms will again vary, at least to some extent, dependent on the target's particular circumstances. (For instance, how emotionally robust they are, or how many good friendships they have, at the point at which they are blamed). Establishing how harmful our blame is – and whether it is nontrivially so – will require us to attend to this kind of contextual variation.

With regards to Blameworthiness-Component the contrast between *SMRT: Retributivism* and *SMRT: Claim Forfeiture* is clear. The former understands the relevant blameworthiness in terms of a specific kind of positive reason (non-instrumentally good harm). The latter understands it in terms of a forfeited claim, plus a *non*-specified positive reason to act in a way which would otherwise violate that claim.

3.3.2. Menges' Desiderata

How does Menges come to deem his view the *correct* understanding of basic desert? He employs distinct (though complementary) desiderata for the relevant blameworthiness and blame-response respectively.[143] (So, we can readily interpret these as desiderata for Blameworthiness-Component and Blame-Component in SMRT). There are three desiderata for each, which I now briefly summarise.

Menges' (2023: 617-621) first two blameworthiness desiderata describe conceptual entailments which he thinks obtain in any appraisal of deservingness – and thus will necessarily obtain in the particular kind of deservingness in question (basic desert). First, the relevant blameworthiness (since it is a form of deservingness) should account for how the blameworthy agent lacks any standing to demand not to be subject to blame which they basically deserve (nor demand an expression of regret or compensation if they have been so blamed). I will call this the *No Standing* desideratum. Second, it should explain how there is something normatively positive about blaming when basically deserved – something which speaks in favour of such responses. Call this the *Positive Reason* desideratum. The third desideratum responds to something else: the character of the longstanding debate between moral responsibility skeptics and defenders. This debate, as Menges notes, is characterised by arguments concerning "luck, determinism, control or God" and their implications for free will and moral responsibility. So, this desideratum asks that the relevant blameworthiness makes sense of this debate: its controversial commitments should

[143] Menges does not describe the first set of desiderata as pertaining to blameworthiness, but rather to "an important entailment of a [basic desert] claim, namely the appraisal or evaluation that is part of it" (2023: 616). He is therefore seeking to capture the *normative significance* which basic desert claims (properly understood) attribute to sheer agential features – how it is that such features are appraised as generating a normative reason. I suggest it is reasonable to interpret this as equivalent to capturing the properties of the relevant kind of blameworthiness, at least given how I am understanding that concept.

concern the same terrain which has been contested by the parties of the debate. Call this the *Same Debate* desideratum.

What about Menges' three desiderata for basically deserved *blame*? The first two take their lead from practical implications which basic desert skeptics – principally Pereboom and Caruso – have explicitly ascribed to their skepticism. First, that the relevant blame is one such that skepticism about its appropriateness would imply we need to make significant *revisions* to both our interpersonal practices and our public institutions**.** Second, the relevant blame should tally with skeptics' claims that it is *unjust* if we subject the non-blameworthy to it. Menges (2022: 2-4) terms these the "revisionary constraint" and "justice constraint" respectively. The third desideratum is different in kind: it should not be trivially true that basically deserved blame is sometimes appropriate. This, which Menges terms a "neutrality constraint", is a version of a principle of charity. We should charitably assume that whatever the blame response in question is, the skeptics have reasonable grounds to be skeptical about its appropriateness.[144]

I summarise these desiderata with a view to criticising how Menges makes use of them. Before critiquing Menges' arguments however, it will help to get clear on what I am *not* disputing. First, I do not dispute that both sets of desiderata represent plausibly sourced standards against which candidates for Blameworthiness-Component and Blame-Component can be measured. So, with regards to the blameworthiness desiderata, grant Menges' conceptual claim that all genuine appraisals of deservingness will satisfy both *No Standing* and *Positive Reason*.[145] And

[144] Although Menges does not say so, neutrality presumably demands that we are also charitable in the other direction. So, for any plausible Blame-Component candidate, *non*-skepticism about its appropriateness should likewise be a *prima facie* reasonable position (or: it should not be trivially true that such blame is never appropriate).

[145] Although a clarification seems necessary with *No Standing*. Where one deserves a certain treatment, one surely *can* reasonably insist they do not receive that treatment. For instance, assume Frank deserves to be punished, and say that we correctly ground Frank's deservingness of punishment in his poorly exercised reasons-responsiveness. Frank surely retains the standing to demand not to receive his punishment on grounds *other than those which pertain to his deservingness*, if he considers such grounds exist. His demand cannot take the form: 'Do not punish me – consider how well my reasons-responsiveness was exercised!'. But he could yet point to *other* countervailing reasons which exist – for instance the immense distress it will cause his mother – reasons which, let's assume, outweigh the reason of desert. In which case, he retains the standing to demand not to be blamed in spite of his deservingness – or at least to make such demands with reference to the demands of morality. (It is perhaps harder to see how Frank would retain standing to demand apology or personal compensation for receiving his deserved blame – but again, perhaps some form of regret or compensation is due

there is indeed well-established metaphysical terrain over which the skepticism debate has been fought, centring upon whether our world can accommodate the freedom or control requisite for free will and moral responsibility.[146] The *Same Debate* desideratum captures this. Likewise with regards to the three Blame-Component desiderata: skeptics including Pereboom and Caruso evidently *have* taken their arguments to have revisionary implications and to bear upon matters of justice and fairness.[147] And principles of charitable interpretation, as embodied by Menges' neutrality constraint, are very well-established.

Neither do I contest Menges' following arguments, briefly summarised, against several possible candidates for Blameworthiness-Component and Blame-Component, on the grounds that they fail to satisfy at least one of the relevant desiderata (2023: 624-627). He rejects understanding Blameworthiness-Component in terms of *fittingness* in the way that Dana Nelkin (2016) understands desert.[148] For Menges, blame's being fitting does not cancel one's standing to demand not to be blamed or for apology or compensation post-blame. Nor does it provide a positive moral reason to issue blame (something which Nelkin's account explicitly accepts). So, a fittingness account fails to meet both the *No Standing* and *Positive Reason* desiderata. Menges also rejects understanding Blameworthiness-Component in terms of *justified* blame. He proposes a case where blame is justified yet its recipient maintains standing to demand compensation or an expression of regret. So, blameworthiness qua justified blame does not meet the *No Standing* desideratum.[149]

With regards to Blame-Component, Menges (2022: 4-5) dismisses as suitable candidates each of the 'constitutive' senses of blame I discussed Chapter 1. So, basically deserved blame cannot comprise merely a judgment (of ill will or fault); nor the combination of a judgment with some non-judgmental response such as a desire,

---

having failed to meet the demands of morality). Therefore, I suggest Menges' *No Standing* desideratum is more precisely articulated as lacking the standing to demand not to be blamed *with reference to those facts which pertain to their basic desert*.

[146] Hence why my analysis of blameworthiness in Chapter 1 foregrounds matters of (agential) control.

[147] Vilhauer (2012) also provides a useful overview of how free will skepticism is revisionary.

[148] McCormick (2022) has also recently argued that basic desert ought to be understood in terms of fittingness (of reactive attitudes).

[149] Note I have already argued it is mistaken to attribute this view to Caruso and Morris, as Menges does - see Footnote 134.

or a modification of relationships of the sort Scanlon describes. He also rejects those accounts which understand blame in terms of a function, whether that be protest, communication, signalling commitment to norms or initiating conversation. For Menges, each of the above violate the neutrality constraint – a skepticism which held that such kinds of blame could never be warranted on the basis of sheer agential features would seem trivially false. (Although, as Menges acknowledges, none of these has ever in fact been proposed as the type of blame implicated in basic desert, or more generally as that which is the target of free will or responsibility skeptics. This might suggest less pressing need to rebut these as candidates for Blame-Component).

Menges (2022: 6-8) also rejects understanding Blame-Component in terms of a reactive emotion (such as resentment) or the expression of such an emotion. He thinks such a proposal would violate the justice constraint since reactive emotions, and their outward expression, are not necessarily responses which require moral justification. I am unconvinced here. It seems very plausible to me that such responses might, at least *sometimes*, amount to an injustice. For example, expressions of vituperative resentment directed at a visibly fragile individual might well seem unjust where they are not blameworthy. Such a response would meet the justice constraint. Menges recognises this possibility too, but he claims that "skeptics need an account of blame which makes sense of the idea that *every* instance of this kind of blame [expressed emotional blame] is unjust when the target is not blameworthy and when other things are equal" (2022: 8, italics added). But I do not see why, even if we accept the justice constraint, skeptics need to identify Blame-Component with a blame-type wherein *every* token of that blame-type amounts to an injustice. Indeed, Menges' own *Claim Forfeiture* account does not meet such a stringent requirement. It identifies Blame-Component not with any specific blame-type, but instead with whichever type of blame non-trivially negatively affects its recipient and thus violates a standing claim. His account therefore naturally meets the justice constraint. But, as discussed, there is inherent variability in which of our responses will amount to a claim violation, and no guarantee that *all* tokens of any specific blame-type will amount to a claim violation, even if some do and are thus in scope. Therefore, contra Menges, expressed emotional blame (for example) could yet – at least sometimes – satisfy the justice constraint and be in scope for *SMRT: Claim Forfeiture*.

Having dismissed each of these various possibilities, Menges argues that the proposals for Blameworthiness-Component and Blame-Component which comprise *SMRT: Claim Forfeiture* do, in contrast, satisfy each of their respective desiderata. This too I will not dispute.[150] So, I do not challenge his unfavourable assessment of these rival views against the desiderata, nor the way he favourably applies them to his own account. What I *will* challenge are his arguments that *SMRT: Retributivism* cannot meet the desiderata. (Then, in Section 3.3.5, I raise doubts about how useful Menges' desiderata in fact are for identifying SMRT, the theory of moral responsibility at stake in basic desert skepticism).

Here it is worth highlighting a final feature of Menges' account. He diagnoses vagueness in the skeptics' description of basic desert, as do I. But he does so having only considered Pereboom's original *Standard Characterisation*, and not the updated version. Nor does he consider Pereboom's recent commitment to non-instrumentally good harm as what grounds the relevant reason. He also does not appear to reckon with Pereboom's (and Caruso's) more consistently stated view that it is singularly moral responsibility as *accountability* which is the object of his skepticism. All of which means he misses the (somewhat) more precisely drawn, retributivist conception of basic desert which I attribute to the skeptics (albeit one with residual vagueness in its conceptualisation of blame and praise).

So Menges' arguments carry more significance than he recognises. Amongst the various possible rival interpretations of basic desert which he rejects is the very interpretation which – even if he himself does not recognise it as such – the skeptics *themselves* favour. If his arguments here succeed, this will mean that Pereboom and Caruso are not merely vague in their characterisation of basic desert, they are *incorrect*. The next two sections aim to resist these arguments.

[150]None of this should not be taken as a full-throated endorsement of Menges' arguments – either those which conclude these other candidate views fail to meet the desiderata, or those which conclude that his own account meets them all. Rather, I cursorily mention and am willing to accept them without criticism, because ultimately my arguments in Section 3.3.5 reveal these desiderata are not in fact apt for determining SMRT, and so it is not necessary to comprehensively assess all possible contenders for SMRT against them.

3.3.3. Defending *SMRT: Retributivism* – Blameworthiness as Non-Instrumentally Good Harm

SMRT: Retributivism interprets Blameworthiness-Component in terms of non-instrumentally good harm, in line with McKenna's account of basic desert.[151] Menges objects to this interpretation. On his view basic desert entails the existence of *some* positive reason to blame, and he leaves it open how this reason might be specified. But he is clear that this reason cannot be specified in terms of the non-instrumental goodness of blame's harm. This is because such a proposal purportedly fails to satisfy two out of his three blameworthiness desiderata for basic desert. I disagree on both counts.

Non-Instrumental Goodness and the *No Standing* desideratum

First, Menges argues that the McKenna-style view cannot satisfy *No Standing*. He uses the following example in support of this claim:

> It may be non-instrumentally good for you and from the point of view of the universe that you and Dumbledore become good friends such that there is reason that favors your becoming friends with Dumbledore. Imagine that I can magically make it the case that you and Dumbledore become good friends. You surely have the standing to demand from me that I don't do it if you do not want to be his friend. If I, nonetheless, magically make you become friends, I owe you an expression of regret, an apology, and perhaps (symbolic) compensation (2023: 622).

This example demonstrates that if action or response X realises non-instrumental goodness, this is compatible (at least sometimes) with the target of X preserving their standing to demand not to be subject to X, and to expect remorse, apology or compensation for having been subjected to X. Whereas their deservingness of X would be incompatible with their preserving this standing (as per the *No Standing* desideratum).

[151] At least, Pereboom understands the relevant blameworthiness in this way (whereas Caruso is less committal). Also, whilst I remain focused here on blameworthiness, *SMRT: Retributivism* understands praiseworthiness in parallel fashion, in terms of non-instrumentally good beneficial praise.

These findings seem very plausible. Menges thinks they imply that views such as McKenna's and *SMRT: Retributivism* cannot satisfy the *No Standing* desideratum. This is where I disagree. These views are not wedded to the (much less plausible) claim that *all* actions or responses which realise non-instrumental goodness are incompatible with preserving the standing in question. They need only insist on this incompatibility for a distinct subset of those responses – namely, those which are basically deserved. To see this, consider how McKenna construes reasons of desert:

> Desert supplies reasons to *discount* or *silence* – and not merely outweigh – moral considerations that otherwise apply. Grant that morality demands of us that when interacting with the moral community we do not, as Strawson himself expressed it, "acquiesce in the infliction of suffering" … at least not unless it can be justified for instrumental reasons. Reasons of desert supply reasons to discount or silence that sort of moral reason and thus support forms of treatment otherwise opposed to morality's counsel (2024: 158).

For McKenna then, basic desert does not merely supply a new reason, grounded in non-instrumental goodness, to respond with harmful blame. Because it is a species of desert, it also entails the *discounting* or weakening of certain countervailing reasons which would otherwise apply. This idea, I suggest, could naturally be translated into Menges' terms of claim forfeiture. The discounted reasons are those grounded in our claims against being non-trivially harmed (treatment which is typically 'opposed to morality's counsel'). Because the blameworthy agent has forfeited these claims, they thereby relinquish their standing to appeal to these particular reasons.

McKenna describes reasons which have this discounting function as *exclusionary* reasons. (2024: 158). Such reasons do not merely outweigh certain countervailing reasons, they exclude them, so that those countervailing reasons should no longer carry any weight in our deliberations about what to do.[152] It is this exclusionary feature of basic desert – rather than the non-instrumental goodness of the blame – which would explain why the blameworthy agent lacks standing to demand restraint or redress. So, it is this feature which explains how McKenna's view can satisfy *No Standing*. The problem with the Dumbledore example is that it is disanalogous in this respect. We

[152] See Wallace (2019) and Raz (1977) for discussion of the notion of exclusionary reasons (with particular focus on how promises generate such reasons).

are given no grounds to think that the reason which favours your becoming friends with Dumbledore is *exclusionary* in the way that McKenna claims desert-based reasons are.[153] In the example, the universe is apparently simply configured such that non-instrumental goodness accrues upon your gaining Dumbledore's friendship. Any reasons against this befriending seemingly remain in place, undiscounted, as does your standing to appeal to them (as Menges points out). Whereas in the case of basic desert, the opportunity for non-instrumental goodness to be realised has distinctive origins: the deployment of sheer agential features in negative moral performance.[154] These origins are essential in explaining that feature which is absent in the Dumbledore case: the exclusion of, and loss of standing to appeal to, certain other countervailing reasons.

Therefore, interpreting Blameworthiness-Component in terms of non-instrumentally good harm – as *SMRT: Retributivism* does, echoing McKenna's account – may yet be compatible with satisfying *No Standing*. The notion of exclusionary reasons makes room for this possibility.

Non-Instrumental Goodness and the *Same Debate* desideratum

Menges argues that views which understand Blameworthiness-Component in terms of non-instrumentally good harm also fall foul of the *Same Debate* desideratum. Why? Because a common view amongst philosophers is that harming is *never* non-instrumentally good.[155] So, the objection goes, if we interpret Blameworthiness-Component in terms of non-instrumentally good harm then those who share the commitment that harming is never non-instrumentally good are effectively straitjacketed into skepticism about basic desert. Menges thinks that, if the *Same Debate* desideratum is to be met, this cannot be the appropriate basis for skepticism:

> [S]kepticism is typically presented and plausibly thought of as the conclusion of substantial arguments about luck, determinism, control, or God – not as the view

[153] It is curious that Menges ignores this crucial exclusionary feature of the McKenna-style account, given that his own proposal for how we should understand Blameworthiness-Component has an explicitly exclusionary dimension. It involves the exclusion of some reasons against being blamed, owing to forfeited claims, as well as the creation of new reasons favouring blame.

[154] McKenna, as we have seen, argues that reasons-responsiveness is the relevant agential feature.

[155] Those notably explicit in holding this view include Parfit 2011; Scanlon 2013; Smith 2019; List in List, Caruso, and Clark 2020.

> one starts with as soon as one understands the concepts at issue and holds the independent view that harming is never non-instrumentally good (2023: 624).[156]

In response, I first note that not *all* authors subscribe to the impossibility of non-instrumentally good harm. In Section 3.2 we met some retributivists, for instance, for whom the existence of non-instrumentally good harm is an indispensable part of their view. So, the impossibility of non-instrumentally good harm remains a reasonably open question. Even those who believe that non-instrumentally good harm is impossible, then, may yet seek alternative grounds to support their skepticism about McKenna-style basic desert.

Indeed, this is what the basic desert skeptics have done, offering alternative counterarguments which use the concepts which Menges correctly identifies as integral to the skeptical debate. For instance, take the manipulation argument, one of the most prominent in Pereboom and Caruso's arsenal (Pereboom 1995; 2001: 110-117; 2014: 71-103; Caruso 2021: 35-108; Caruso & Pereboom 2022: 25-30). This argument aims to show that determinism precludes the kind of agential control which compatibilist theories of moral responsibility (of which reasons-responsiveness theories such as McKenna's are a paradigmatic example) appeal to. In my view, work remains to be done to appreciate the full implications of the manipulation argument and establish exactly which kind(s) of blameworthiness it undermines. This is work I undertake in the final chapter. But we have already seen that Pereboom explicitly understands his skepticism about basic desert – in which the manipulation argument is a central plank – to target the notion of non-instrumentally good harmful blame (and beneficial praise).[157] So, it seems clear that blameworthiness understood in terms of non-instrumentally good harm lies at the heart of the skeptical debate concerning free will and basic desert moral responsibility.

More generally, any plausible *Same Debate* desideratum surely does not require that for any credible basic desert proposal there are *no* credible counterarguments available other than those which invoke the concepts of luck, determinism, control or God.

[156] Vicens (2022: 359) voices what I take to be the same concern: "[w]e do not want to rule out the possibility that anyone could basically deserve any kind of responses to their behaviour just because we are opposed to the idea that it could ever be, say, intrinsically good to make someone suffer".

[157] Indeed McKenna (2008, 2014) and Pereboom (2008) have directly debated the implications of the manipulation argument for the McKenna-style view.

Rather, what is required is that the basic desert proposal invites counterarguments which draw on such concepts.[158] McKenna's proposal evidently does this, and thus I suggest it comfortably satisfies *Same Debate*. If we have independent grounds for rejecting the possibility of non-instrumentally good harm, these simply provide *extra* reason to be skeptical about a McKenna-style proposal. Pereboom himself describes this as an "additional objection" which might be levelled against basic desert (2021b: 31).

Therefore *SMRT: Retributivism* satisfies the two Blameworthiness-Component desiderata which Menges claims it cannot. The result is a partial reconciliation between *SMRT: Retributivism* and *SMRT: Claim Forfeiture*. The latter account, we saw in Section 3.3.1, conceives Blameworthiness-Component as a composite of two things: the forfeiture of a claim against non-trivial harm, and a reason which favours blame in the form of such harm. It leaves open whether any such favouring reason exists and if so, what it might be.[159] I submit that a reason grounded in non-instrumentally good harm, in satisfying his Blameworthiness-Desiderata, should be accepted by Menges as one which plausibly fits the bill and thus could complement his *Claim Forfeiture* account.[160] (By the same token, Menges' objections concerning non-instrumental goodness are not adequate grounds for rejecting McKenna's account as a genuine theory of basic desert. I instead see the problem for that account as the particular non-instrumental goods it identifies cannot be squared with the sheer agential features condition. Although I have not set out this argument).

[158] This means the *Same Debate* desideratum – at least this more plausible construal of it – is actually very easy to meet. *All* theories of moral responsibility – at least as I have suggested we understood them in Chapter 1 – necessarily involve commitments about the compatibility between agential control and determinism (or indeterminism, in the case of incompatibilist theories). Counterarguments which concern the relationship between these same concepts will therefore be apposite for all such theories.

[159] Hence the *Claim Forfeiture* account straightforwardly satisfies the *Positive Reason* desideratum (by stipulating that basic desert entails the existence of a positive reason to blame). But it does so somewhat cheaply, given it makes no further commitments as to the properties of this positive reason.

[160] To be clear, my claims here – like Menges' opposing claims – are solely conceptual, rather than normative. I claim there are conceptual resources available which allow for understanding Blameworthiness-Component in terms of non-instrumentally good harm without necessarily prohibiting the *No Standing* and *Same Debate* desiderata from being satisfied. This is in no way to defend the normative claim that blameworthiness of this sort in fact exists, nor that this fact in turn generates exclusionary reasons, in the way that McKenna argues. (In the previous chapter I briefly sketched why we might reject this claim with regards to the particular non-instrumental reasons which McKenna describes. In the next chapter I reject it more universally).

The Unsuitability of Communicative Reasons

In fact, I think that a reason grounded in non-instrumentally good harm represents a more promising candidate for the relevant favouring reason than Menges' preferred option. Again, the *Claim Forfeiture* account – wishing to remain a strictly conceptual and non-normative thesis – is agnostic about the existence of reasons favouring basically deserved blame. But Menges does briefly propose one kind of reason which he thinks has normative merit and could be bolted onto his conceptual account. The reason is described in terms of blame's *communicative* value:

> We often have reason to communicate that we want or demand that the blameworthy acknowledge their fault, dissociate from it, and, perhaps, offer compensation. Ideally, the blameworthy would publicly and sincerely accept that they have wronged someone, disrespected a moral value, or violated a moral norm, that they should not have done it, and promise that they will not do it again. Blaming the blameworthy communicates that we want or demand that they acknowledge their fault in such a way (2023: 630).

As Menges notes, his endorsement of such reasons for blaming puts his account into alliance with established communicative theories of moral responsibility.[161] Such theories identify similar valuable roles which blame can fulfil as a communicative act. The problem, I think, is that communicative theories are unable to accommodate a core feature of basic desert: blameworthiness generated out of sheer agential features. Whatever else basic desert involves, it *inarguably* has this feature, one which all disputants explicitly accept.

Menges seems not to see this problem, as he suggests that sheer agential features deployed in negative moral performance are sufficient for grounding such communicative reasons. In his words, "people's objectionable behavior combined with some facts about how the agents were when they behaved this way – such as what they knew and what control they had" are what provide the grounding (2023: 630). But this cannot be the whole story. Consider the valuable outcomes which Menges

[161] Those who have proposed communicative theories of moral responsibility – or at least theories in which blame plays an important communicative role – include Darwall (2006), Fricker (2016), Shoemaker (2015: 64-86), McKenna (2013), and Macnamara (2015). These communicative theories however do not typically emphasise the *harmfulness* of the blame they endorse, whereas this is an essential feature of Menges' account of basically deserved blame.

describes – gaining the blamed person's acknowledgment and dissociation from wrongdoing, and their provision of compensation and assurance of reformed behaviour. These are outcomes which blame secures instrumentally via the recipient's uptake of the blame's communicated message. In order for this to generate a reason to blame, we need to know more than that the recipient's sheer agential features were deployed in objectionable moral performance. We need an additional normative consideration about the desirability of these outcomes (as well as a consideration about the instrumental efficacy of the blame in securing them). Absent these additional considerations, we would lack this communicative reason to blame. So, communicative reasons stray beyond the parameters of basic desert because they concern what blame is *instrumentally* able to achieve. As discussed in Section 2.4.2, reasons favouring blame must instead be non-instrumental if they are to be grounded in sheer agential features, which is the hallmark of basic desert.

Indeed, this explains why Pereboom and Caruso themselves are happy to *endorse* communicative reasons to blame. Pereboom has consistently advocated blame used as a tool to communicate moral protest and achieve outcomes akin to those championed by Menges. Pereboom describes these outcomes as "protection from harmful immoral behavior, reconciliation with those who have done one wrong, moral formation of those disposed to do wrong, and retention of integrity of victims of wrongdoing" (2021b: 36).[162] It is understandable why the skeptics deem such communicative reasons to be unthreatened by their skepticism about blameworthiness in the basic desert sense. These are not reasons generated by sheer agential features.

3.3.4. Defending *SMRT: Retributivism* – Blame as Intentional Harm

What about Blame-Component? *SMRT: Retributivism* specifies the relevant blame in basic desert as intentional harm.[163] Menges argues against this 'punishment' conception, a position he (rightly) associates with Caruso – although we have established that Pereboom likewise interprets Blame-Component in punitive terms. Recall too that since the intentional harms in question are inflicted on the basis of sheer

[162] See also Pereboom 1995, 2001, 2014, 2021b; Caruso 2021; Caruso in Dennett and Caruso 2021; Caruso and Pereboom 2022.

[163] I will ignore how *SMRT: Retributivism* specifies Praiseworthiness-Component (in terms of intentional benefitting), as I remain focused on responding to Menges' arguments which exclusively concern blame.

agential features (deployed in negative moral performance), they can be appropriately understood as *retributively* punitive blame.

Menges argues that, whilst such blame easily meets the neutrality and justice constraints, it has problems meeting his revisionary constraint on basic desert. Were skepticism about basic desert in this sense to win out, and it be established that such retributively punitive responses were unjustified, Menges thinks this would *not* imply any significant revision to our interpersonal practices because our practices do not involve any serious commitment to intentional harming.[164] In support of this claim, Menges cites only his personal experience: "I hope and believe that this intention is rarely part of my everyday life" (2022: 6). But this is very scant evidence from which to draw conclusions about the prevalence of interpersonal retributivism, against which we can adduce several empirical studies which indicate a willingness amongst the folk to punish for retributive reasons.[165] These studies suggest that skepticism about retributive blaming would in fact be significantly revisionary for our interpersonal practices. At the very least, then, the revisionary implications of skepticism about retributive blame for our interpersonal practices remains an open empirical question.

Furthermore, Menges readily admits that skepticism about the appropriateness of retributive harming would be revisionary for our *institutional* practices. This is because many instances of criminal punishment are justified, at least partly, in retributive terms (I return to this, with some examples, in Section 4.7 of Chapter 4). This accounts for why skeptics, in considering the practical implications of their skepticism, have mostly focused on prescriptions for criminal reform.[166]

[164] There is a tension in Menges' arguments here. Menges claims that intentional harms (punishment) are not the correct way to understand Blame-Component, because this fails to meet the revisionary constraint (2022: 6). But he also explicitly includes intentional harming amongst his examples of the kind of blame implicated in basic desert, according to his account (2022: 11). The rationale for doing so is clear: intentional harms are very obviously a kind of response which we typically take to violate interpersonal claims against non-trivial harm. Charitably, I suggest we understand Menges' position as follows: intentional harms *would* be a kind of blame which can be basically deserved were intentional harming sufficiently endemic to our interpersonal practices that it would be revisionary to discover they are unjustified. But because Menges thinks such a discovery would not be revisionary, he concludes such blame is not what is stake in basic desert.

[165] Notable studies include McFatter (1978, 1982); Kahneman et al (1998); Darley et al. (2000); Carlsmith et al. (2002); Carlsmith (2006, 2008); Carlsmith and Darley (2008), Nadelhoffer et al (2013).

[166] For instance see Shaw, Pereboom and Caruso (2019), and also the recent work Pereboom and Caruso have done to develop a Public Health-Quarantine Model for criminal justice (Pereboom 2013, 2021b: 78-102; Caruso 2021: 184-328, Caruso and Pereboom 2022: 55-60).

Therefore it is not at all clear that interpreting Blame-Component in terms of intentional retributive harm flouts the revisionary constraint, as Menges claims it does. To the extent that the revisionary constraint is satisfied, we have a second point of reconciliation between *SMRT: Retributivism* and *SMRT: Claim Forfeiture*. Blame in the form of intentional harm is, at least very often, a kind of response which violates interpersonal claims against non-trivial harm. Therefore, since it also satisfies Menges' blame desiderata, it is the sort of blame which the *Claim Forfeiture* account ought to recognise can (conceptually-speaking) be basically deserved.

In sum, *SMRT: Retributivism* can withstand Menges' criticisms. Its two central commitments – specifying Blameworthiness-Component in terms of the non-instrumental goodness of harmful blame, and Blame-Component in terms of intentional harm – are able to meet the respective desiderata. In fact, Menges should endorse these as *examples* of the kind of blame and blameworthiness which his own *Claim Forfeiture* takes to be implicated in basic desert.

3.3.5. The Normative Inadequacy of Menges' Desiderata

I have claimed that *SMRT: Retributivism* – the interpretation of basic desert held by the skeptics – can satisfy Menges' desiderata. But what is the significance of this? Menges thinks the desiderata can be used to determine the kind of blame and blameworthiness which basic desert skeptics ought to be skeptical about. But I have doubts about how suited the desiderata really are to this task.

I think it is important here to distinguish between two different areas of uncertainty which seem to get blurred in Menges' analysis. The first is the ambiguity in what the skeptics have meant by basic desert. What precisely is this kind of moral responsibility which they take their skeptical arguments to undermine? Or equivalently, how do *they* conceive SMRT? We might call this the exegetical uncertainty. The second area of uncertainty concerns the kind of moral responsibility which their skeptical arguments in fact weigh against. Call this the normative uncertainty: what theory of moral responsibility should we be skeptical about? What is the *correct* characterisation of SMRT? We would expect, and certainly hope, that the answer to both sets of questions will coincide, so that the skeptics' characterisation of basic desert matches the kind of moral responsibility which their skeptical arguments in fact undermine. I ultimately aim (in Chapter 4) to demonstrate how this is not straightforwardly the case. For now, I aim to show how Menges' desiderata are of limited use in resolving either of these two uncertainties.

Take the exegetical uncertainty. The desiderata might appear useful as tools for identifying the version of SMRT which the skeptics have ambiguously characterised. They allow us to test for forms of blameworthiness and blame which have certain qualities that the skeptics have associated with basic desert. So, a blameworthiness which qualifies as a *bona fide* case of deservingness (*No Standing* and *Positive Reason*), about which skepticism has revisionary implications which are a matter of justice (revisionary and justice constraints). They also ensure it is a form of blameworthiness which invites counterarguments of the relevant conceptual tenor (*Same Debate*), whilst not uncharitably saddling the skeptics with an implausible view (neutrality constraint).

However, there are good reasons to doubt the usefulness of the desiderata for this exegetical task. First, I have argued that *SMRT: Retributivism* and *SMRT: Claim Forfeiture* are *both* able to satisfy the desiderata, in which case these are not criteria which vindicate one of these accounts over the other. More significantly still, in Section 3.2 I addressed the exegetical uncertainty. I showed that, by taking account of the full range of descriptive claims which Pereboom and Caruso make about basic desert, we can piece together a reasonably clear picture of how they interpret SMRT (albeit one with residual vagueness in how they conceptualise blame and praise). The view we arrived at is *SMRT: Retributivism*. The skeptics' interpretation of SMRT is therefore much less ambiguous than Menges appreciates.[167] We need not rely on the desiderata to establish it.[168]

Having rendered the skeptics' position more precisely, there is still room to ask whether this, or some other view, reflects the *correct* interpretation of the relevant skeptical considerations. This – what I called the *normative* uncertainty – is what I seek to resolve in the rest of the thesis. But here the desiderata are arguably even less

[167] As discussed, Menges refers only to Pereboom's original (and more decidedly ambiguous) *Standard Characterisation* of basic desert.

[168] Establishing this less ambiguous version of the skeptics' account would also seem to render the neutrality constraint somewhat redundant. If there is less room for interpretation about how the skeptics conceive of SMRT, there is less scope for the charitable interpretation of this view which the neutrality constraint aims to ensure.

instructive. At best, they might serve as useful *proxies* for identifying the correct version of SMRT.[169] But this is only if we assume that the skeptics are accurate in taking their skeptical arguments to specifically undermine a kind of deservingness, and to have revisionary implications which are a matter of justice. However, these assumptions might be misplaced. Clearly the only determinative way to resolve the normative uncertainty is to interrogate the skeptical arguments directly and establish which theory of moral responsibility (if any) they in fact tell against. Without consulting the skeptical arguments in this way, Menges' desiderata are an inadequate guide.[170]

So, I claim the desiderata are superfluous for resolving the exegetical uncertainty, and inadequate for resolving the normative uncertainty. The desiderata may well *track* both the conception of SMRT held by the skeptics and the correct conception of SMRT (at least, nothing I have argued implies they do not). It is just that they are inadequate for settling either of these things. But if I do not take the desiderata to be probative in either case, why devote Sections 4.3.3 and 4.3.4 to demonstrating how *SMRT: Retributivism* (contra Menges) satisfies them? This is still worthwhile as it would be significant if *SMRT: Retributivism* failed to satisfy the desiderata. It would imply inconsistencies in the skeptics' claims and cast new doubt on what their settled position is. In confirming that their conception of basic desert satisfies the desiderata, we *corroborate* their accompanying claims about its qualities. In particular, that it is appropriately labelled as a desert thesis, and that skepticism about it is genuinely revisionary and a matter of justice. But having shored up the skeptics' position in this

[169] In fairness, Menges describes these standards as desiderata, so perhaps he is happy to see them as proxies rather than determinative criteria. Still, he uses them to derive the kind of moral responsibility which is at stake in basic desert skepticism – whereas I claim a much better test is available.

[170] The *Same Debate* desideratum is slightly different, as this is itself derived directly from the conceptual make-up of the arguments offered by basic desert skeptics. Any theory of moral responsibility which these skeptical arguments are effective against – and which is thereby shown to be the *correct* interpretation of SMRT – will necessarily be a theory which makes contestable claims concerning matters of control and (in)determinism (thus satisfying *Same Debate*). But as mentioned in Section 4.3.3, this is a very low bar to meet, since *all* theories of moral responsibility necessarily involve claims about the relationship between control and (in)determinism, or so I argue. Certainly this is a bar which we have seen *SMRT: Retributivism*, as well as *SMRT: Claim Forfeiture*, can easily meet. So, unlike the other desiderata, we can be certain that they will necessarily satisfy *Same Debate*. But this does little to pinpoint what the correct interpretation of SMRT in fact is. Again, clearly, it is only by closely interrogating what the skeptical arguments are able to establish (rather than merely registering the concepts which they invoke) which can determine this.

way, we must next interrogate their skeptical arguments to establish whether this position accurately reflects what it is which those arguments serve to undermine.

Interestingly, Menges himself proposes we need to revisit the skeptical arguments, in light of the (purported) vindication of the *Claim Forfeiture* account:

> [S]keptics should argue that because of determinism or luck, human misconduct and the way humans are when they perform it never explains why they forfeit claims… From this perspective, manipulation, rollback, replication, and luck arguments try to show that, under certain conditions, agents do not forfeit their claims because of their objectionable conduct and how they were when they performed it. Spelling out the implications of [the *Claim Forfeiture* view] for the arguments for and against skepticism is a job for future research. (2022: 13)

I think this get things the wrong way round. It seems unusual to suggest that the skeptical arguments *should* aim at this – or any other – conclusion. Instead, we ought first to establish precisely which theory of moral responsibility the skeptical arguments undermine, whether this is *SMRT: Claim Forfeiture* or some other. It is the skeptical arguments which will have implications for, and potentially vindicate, the *Claim Forfeiture* view, not the other way round.[171] This is the approach I seek to model in the following chapter.

3.4. A Much Wider Church: *Basic Reason Moral Responsibility*

I have set out two distinct conceptions of basic desert. I have rejected Menges' desiderata-based methodology as inadequate for adjudicating whether these (or any other) conceptions are the correct way to understand the kind of moral responsibility which is challenged by basic desert skeptics (SMRT). Instead, we must consult the skeptical considerations directly. Before doing so, I now sketch a third candidate for SMRT. This can be done briefly, as this proposal – which I will call *Basic Reason* moral responsibility – involves little additional conceptual machinery. Unlike the first

[171] Perhaps one might query my own approach here. Why not begin with the analysis of the skeptical arguments, and *then* use this to establish whatever conception of moral responsibility it is that they serve to discredit? That sequencing may well work, and it would be similar in spirit my approach. But I think it is useful to have these two prominent theories of basic desert on the table first (plus a third option – *SMRT: Basic Reason* – which I set out below). I aim to draw out in Chapter 4 how the skeptics' arguments have particular implications for these theories.

two accounts considered, no one (to my knowledge) has claimed that *this* is the kind of moral responsibility in question.

The *Basic Reason* view understands blameworthiness in terms of *any* normative reason which favours blame. It places no restriction on the normative properties of this reason (for instance, whether it is a reason grounded in the non-instrumental goodness, or rightness, or deservedness, or virtuousness of blaming, or something else). The only relevant restriction is that the facts out of which this reason is generated are limited to sheer agential features deployed in negative moral performance. An agent can also be considered blameworthy in this sense if, rather than generating a new positive reason, their sheer agential features serve to weaken or eliminate counter-vailing reasons *not* to blame them.[172]

What kind of blame are we liable for when we are morally responsible in this *Basic Reason* sense? Equivalently: how would this view specify Blame-Component in SMRT? Very inclusively – *any* action (or omission) is within scope. This theory of moral responsibility therefore strictly concerns normative reasons in favour of *actions* (or omissions): the kinds of consideration which should figure in our deliberations about what to do.[173] Other kinds of blame response – those which are not actions or omissions – are not within its scope. For instance, *Basic Reason* moral responsibility does not bear upon the appropriateness of forming the *judgement* that sheer agential features have been deployed in negative moral performance. Similarly, it is silent on the appropriateness of *emotional* responses in response to sheer agential features (provided that we assume that emotional responses are not actions, as they do not causally originate in an agent's intentions). These are blame responses which might be formed in response to sheer agential features deployed in negative moral performance. But they are not actions or omissions, and so they are not the kinds of

[172] This view understands praiseworthiness in parallel terms: as a normative reason of any description which favours responding with an action. The only difference is that the reason is generated out of sheer agential features deployed in positive, rather than negative, moral performance. But I will continue to focus mostly on blame.

[173] Here I understand *actions* in line with (what has become known as) the standard causalist conception of action. Broadly, according to this conception, actions (and omissions) are events (or absences of events) which bear the right kind of causal relation to certain mental states of the agent. The relevant mental states, at least in more recent accounts, are typically taken to be intentions. We can say that the particular actions of interest here, then, are those which causally originate in the intentions of the agent, where the representational content of these intentions are the relevant reasons of blameworthiness and praiseworthiness. The standard causalist conception of action famously originates in the work of Donald Davidson (1980). See Schlosser (2019a) for an overview.

response for which blameworthiness in the *Basic Reason* sense can provide a favouring normative reason.[174]

So, to qualify as morally responsible – as blameworthy or praiseworthy – in this *Basic Reason* sense, it must simply be the case that one's sheer agential features, deployed in non-neutral moral performance, generate a normative reason of any description to respond with an action (or omission). Or at least that they weaken existing countervailing reasons. Conversely, skepticism about this species of moral responsibility says that sheer agential features being deployed in non-neutral moral performance never alters the balance of reasons in any of these ways.

For illustration, suppose that Janis feels slighted by Julian's offensive (and wrongful) remark. We can imagine a range of possible responses which Janis might have. For instance, a firm demand that Julian apologises; with an imperceptible clench of her fist; by collecting blades of grass; by hiking the Appalachian Trail; by not hiking the Appalachian Trail; by not attending Julian's party later that week. These are *all* responses which fall within the scope of *Basic Reason* moral responsibility. By which I mean, *were* there any kind of normative reason favouring any of these (or any other) actions or omissions, and were this reason generated simply by Julian's sheer agential features as deployed in his wrongful remark, this would be sufficient to render Julian morally responsible in the *Basic Reason* sense. (Julian would also qualify as morally responsible in this sense if existing reasons *against* any actions or omissions, such as these, were weakened or eliminated). But consider instead the resentment which Janis feels towards Julian, or the judgment she forms that Julian has slighted her. These non-action responses are *not* ones which Julian's moral responsibility in the *Basic Reason* sense could render appropriate.

How does *SMRT: Basic Reason* compare with the two other candidates for SMRT discussed in this chapter? First, unlike these other two, this theory of moral responsibility cannot plausibly be described as a theory of basic desert. This is because

[174] Again, I am content to consider all of these response-types (actions, judgments, and emotions) as forms of blame and praise, since I am employing the very thin conception of blame and praise which I set out in Chapter 2. (This conception understands blame and praise to be *any* response which follows the apprehension of agential features deployed in negative and positive moral performance, respectively). If one rejects this conceptual framework, this need not pose a problem for conceptualising the *Basic Reason* position. We can drop the blame-talk, and simply consider *SMRT: Basic Reason* as a theory of moral responsibility which claims only that sheer agential features can sometimes generate reasons in favour of actions.

a core commitment of all desert claims, as discussed in Section 2.4.2, is that the putatively deserved responses are ones which necessarily either favourably or unfavourably affect their recipient. But the kind of blame and praise which fall within the scope of *SMRT: Basic Reason* encompasses more than merely those responses which unfavourably or favourably affect the interests of the recipient. It also includes that large range of actions which do not plausibly affect the blameworthy agent's interests in any way (consider Janis' clenched fist or grass collection). A more apt name for this kind of moral responsibility, then, is *Basic Reason* moral responsibility. This reflects how it understands blameworthiness and praiseworthiness in terms of normative reasons of any description, not limited only to reasons of desert.[175] If it turns out that this is the correct version of SMRT, then, it means that that the arguments offered by basic desert skeptics can be used to undermine more than merely what can plausibly be considered as basic desert moral responsibility. This might seem a surprising result, but I ultimately aim to show it has plausibility.

Another point of comparison worth emphasising is that, in an important sense, these three views are complementary rather than oppositional. Both *SMRT: Basic Reason* and *SMRT: Claim Forfeiture* are non-specific with regards to the normative quality of the positive reason to blame. Whereas, *SMRT: Retributivism* is concerned only with reasons which have a particular normative content: non-instrumental goodness.[176] This, as we have seen, is a kind of reason which can be accommodated by the *Claim Forfeiture* view (despite Menges' claims to the contrary).

In terms of Blame-Component, the three views sit on a spectrum from narrower to wider inclusivity. *SMRT: Retributivism* limits the relevant blame to that which intentionally harms (and praise to that which intentionally benefits). *SMRT: Claim Forfeiture* instead specifies this limit in terms of whether the blame amounts to non-trivial (claim-violating) harm. This represents a broader category of responses, one which includes intentional harms. *SMRT: Basic Reason* broadens things still further, given it encompasses blame in the form of any action. The set of responses identified by *SMRT: Basic Reason* is inclusive of that identified by *SMRT: Claim Forfeiture*,

[175] The name *Basic Reason* also refers to the 'basicness' of those facts – sheer agential features – which generate these reasons, and which is a feature of all three versions of SMRT.

[176] At least, Pereboom (if not Caruso) is explicit that he understands basic desert in terms of non-instrumental goodness.

and this in turn is inclusive of that identified by *SMRT: Retributivism*.[177] We will see this concentric relationship becomes significant when it comes to assessing the implications of the relevant skeptical arguments in the next chapter.

A final point which I consider has practical importance. One feature which is distinctive about *SMRT: Basic Reason* is its relative conceptual simplicity compared to the two basic desert conceptions, which we have seen are *unsettled* in certain ways. This simplicity promises a more straightforward and conclusive story about moral responsibility skepticism, were it to transpire that basic desert skepticism *does* in fact impugn *Basic Reason* moral responsibility. For instance, we could express this skepticism in terms of the inability of sheer agential features to generate *any* normative reason in favour of *any* action (or omission). We would therefore not be obliged to offer an account of praise and blame (which is an unsettled issue in the skeptics' presentation of *SMRT: Retributivism*). Nor, in order to determine the scope of our skepticism, need we identify which of our actions amount to non-trivial, claim-violating harm, and under what conditions (which is an unsettled requirement of *SMRT: Claim Forfeiture*). Nor need we offer a theory of desert, since *all* reasons in favour of action are excluded, not merely reasons of desert.

I would suggest these are very desirable conceptual benefits in the event of *Basic Reason* being vindicated as the correct version of SMRT.[178] But we only get to enjoy them if it transpires that skeptical considerations are in fact effective against this kind of moral responsibility. It is to this that I turn next.

3.5. Concluding Remarks

This chapter has sought to achieve several things. First, I have deciphered more clearly the kind of blame (and praise) which basic desert skeptics profess to be skeptical about, something which is not immediately obvious from their various descriptions of basic desert. We can understand this in terms of intentional harms (and benefits). I also registered how Pereboom, at least, construes the type of blameworthiness at stake in

[177] I ignore here a possibility that *beliefs*, and not just actions, might be wrongful as some have suggested (e.g., Basu 2019). Were this the case it would complicate the 'concentric' picture I describe (since wrongful beliefs – if non-trivially harmful – would be within scope of *SMRT: Claim Forfeiture* but not *SMRT: Basic Reason*).

[178] I will return these considerations in Section 4.7.1 in the following chapter.

basic desert in terms of non-instrumental goodness. Together these commitments culminate in the position *SMRT: Retributivism*.

Next, I set out Menges' rival account of basic desert – the *Claim Forfeiture* view – before defending the *Retributivism* position in light of Menges' desiderata-based critique. I then argued that these desiderata are not in fact a satisfactory tool for the purposes of identifying the kind of moral responsibility put at stake by basic desert skepticism. We must, perhaps unsurprisingly, turn to the skeptical arguments themselves if we are to satisfactorily address this question.

I finished by sketching a third possible answer to the question: *SMRT: Basic Reason*. This describes a kind of moral responsibility which is conceptually streamlined and settled in comparison: it conceives an agent's moral responsibility in terms of there being a normative reason of any description to respond to them with an action of any kind, in virtue of their sheer agential features.

The preponderance of my arguments here have again been, in an important sense, conceptual rather than normative in nature. By which I mean I have not yet offered an argument for which of these three conceptions of moral responsibility, if any, is that which the skeptics' arguments work to undermine. I have just elucidated three possibilities. So, which of these, if any, *does* represent the correct understanding of SMRT? I now turn to consider what the relevant skeptical considerations tell us.

Chapter 4 - Reconstruing Moral Responsibility Skepticism

4.1. Introduction

In this final chapter my attention turns squarely to moral responsibility skepticism. This discussion follows on directly from the themes discussed in the previous chapter. The upshot of that previous discussion was, intuitively enough, that we need to interrogate the skeptical considerations provided by basic desert skeptics directly if we are to determine the kind of moral responsibility at stake in basic desert skepticism (what I have referred to as SMRT). This chapter takes up this challenge. The result of my analysis will be a new, transformational perspective on how such skepticism should be understood.

Here is a fuller outline of the chapter. First, in Section 4.2, I briefly survey the different forms which moral responsibility skepticism has taken and explain my choice to focus on just one skeptical argument. This is the *manipulation argument*, a form of argument well-established in the literature, and I think an especially promising one. This is the argument I draw upon to assess which theory of moral responsibility it is that basic desert skeptics in fact give us reason to doubt (or: which version of SMRT is correct). I focus on the version of the manipulation argument given by Pereboom. In Section 4.3 I summarise Pereboom's argument, and then in Section 4.4 I draw attention to some of its significant conceptual features which have gone underrecognised, and which I argue significantly undermine its force.

In Section 4.5 I look to remedy these conceptual concerns, and simultaneously make good on the promise of the previous chapter, by testing whether the *Retributivism* or the *Claim Forfeiture* interpretation of SMRT is vindicated by the manipulation argument. Does the argument undermine the specific theories of basic desert moral responsibility described by each of these views? I argue that the answer to this question is *yes*, in both cases. In Section 4.6 I then set out a third version of the manipulation argument, this time designed to test the *Basic Reason* interpretation of SMRT. I claim this argument does indeed press us to be skeptical that sheer (compatibilist) agential features can generate *any* reasons for action whatsoever. This means that the *Basic Reason* understanding of SMRT is therefore also vindicated. This implies that the scope of the manipulation argument's conclusion has been underestimated: it recommends skepticism about a much wider range of responses than has previously been understood. I suggest this finding also offers an explanation for how it is that Pereboom's version(s) of the argument, despite the conceptual concerns I raised, can still be effective in eliciting the intuitions it seeks.

In Section 4.7, I consider what these arguments mean for how we should think about basic desert skepticism, arguing that this resolves into a conceptually more straightforward position. In order to grasp what such skepticism entails, we need not decisively determine – in the way that prior presentations of this skepticism imply we must – what such contested concepts as blame, blameworthiness or desert refer to. I see this as a happy, clarifying implication of my analysis. I will also explain what this analysis means for how we assess the two rival conceptions of SMRT, *Retributivism* and *Claim Forfeiture*. These are both views which I still take to capture something significant, but their significance is also now reconstrued. The blame responses which they draw our attention to are distinctive *not* because they are co-extensive with the set of responses undermined by the skeptical considerations of the manipulation argument (hence neither of these views represents the optimal understanding of SMRT). Rather, they are distinctive for a categorically different reason: they are responses which have a particular moral significance, whose propriety (on the basis of sheer agential features) we have particular reason to be interested in establishing or discrediting. I end by briefly outlining some implications which stand even if one rejects my conclusions about *SMRT: Basic Reason*, regarding the methodology I have used.

4.2. Situating The Manipulation Argument Within Moral Responsibility Skepticism

Various kinds of skeptical argument have been put forward by the different skeptics about moral responsibility. I will focus on only one kind – the manipulation argument. Space would not allow for any comprehensive survey of skeptical arguments – but I hope to show how my purposes can be well served by considering the manipulation argument alone.

Set against the wider range of skeptical arguments in the literature we can appreciate how the scope of the manipulation argument is limited in two important senses. First, it is an argument which seeks to establish incompatibilism: the claim that a deterministic world is incompatible with free will and moral responsibility. A deterministic world, put simply, is one in which the past in tandem with the laws of nature necessitates a single unique future (Mele 2006: 3; van Inwagen 1983: 3). The manipulation argument therefore weighs against compatibilist theories of moral responsibility only, and does not purport to argue against libertarian theories (which themselves accept incompatibilism, and are instead premised upon agential features which involve indeterminism of some kind). Different arguments are required to defeat libertarian theories.[179]

Second, the manipulation argument targets only a particular kind of compatibilism: source compatibilism. It aims to show we lack that specific kind of control which this family of compatibilist theories defend as necessary for free will and moral responsibility. Namely, control in the sense of being the appropriate source or author of our actions (rather than in the sense of having the ability or leeway to do otherwise). So, the manipulation argument is an argument in favour of *source* incompatibilism (rather than leeway incompatibilism).[180]

Indeed, most contemporary skeptics can be read as principally targeting the sourcehood form of agential control (Strawson 1986; Waller 1990, 2011; Pereboom 1995, 2001, 2014; Levy 2011; Caruso 2012, 2021). Amongst these authors, prominent skeptical arguments include those which deny we have the necessary sourcehood control because they deem free will to be incompatible with the pervasiveness of luck (Levy 2011). Or those – most notably Galen Strawson's Basic Argument – which

---

[179] As outlined in Chapter 1, libertarian views break down into two main groups, agent-causal and event-causal theories. Skeptical responses to event-causal libertarianism tend to concern whether such views allow for any kind of control necessary for moral responsibility. (Or, translated into my terms, whether these accounts are in fact describing what can be considered as an agential feature – as discussed in Footnote 15). Whereas skepticism about agent-causal libertarianism often concerns its naturalistic plausibility: whether such agential features could even *exist*, given their "panicky metaphysics" as P.F. Strawson might put it, before we even get to the question of their normative significance (1963: 211). Pereboom for instance is open to the possibility that a certain kind of agent-causal agency might allow for basic desert, but he thinks we have very inadequate empirical grounds to imagine such agency actually exists (Pereboom 2014: 50).

[180] Peter van Inwagen's Consequence Argument is probably the most influential argument for the incompatibility of free will and leeway control (e.g., 1975, 1983, 1989). Another hugely significant intervention in that debate was Frankfurt's *Alternate Possibilities and Moral Responsibility* (1969), which argued that the ability to do otherwise is not in fact required for free will and moral responsibility. As mentioned in Section 1.3.3, skeptics have since mostly focused on the issue of sourcehood control and, in examining the manipulation argument, this is also where my attention is.

argue that that free will and moral responsibility (at least of an 'ultimate' sort) are incoherent concepts, since they impossibly require agents to create their own motives and character (Strawson 1994; Sommers 2007).[181]

Basic desert skeptics Caruso and Pereboom (2022: 24) in fact subscribe to *hard* incompatibilism. This is the view that *both* determinism, and indeterminism (at least the kind which might actually be present in our world), are incompatible with basic desert moral responsibility. So they hold that whichever way it turns out our world, and in particular our agential features, are constituted – whether deterministic or indeterministic – basic desert is incompatible with that reality. Their skepticism is therefore global in its scope – it does not rely on the truth of determinism. Many contemporary skeptics – including Galen Strawson (1986), Smilansky (2000), Levy (2011), Double (1991), and Waller (2011) adopt a similar position. Whilst I am also sympathetic to such a global skepticism, my focus will remain on the skepticism of (sourcehood) compatibilism which the manipulation argument specifically aims to motivate.

In analysing the manipulation argument I am therefore only considering one branch of moral responsibility skepticism out of several, leaving aside arguments mounted against libertarian incompatibilist accounts, and (within anti-compatibilism) arguments mounted against leeway compatibilism. This narrowed scope means that any version of SMRT which we derive from the manipulation argument will likewise be restricted – it will concern a theory of moral responsibility premised on sheer *compatibilist* agential features.

But despite this restriction, the manipulation argument claims to be remarkably potent, since it aims to debunk *all* source compatibilist theories, by revealing how *any* proposed compatibilist agential feature will be inadequate for basic desert moral responsibility. If successful it therefore forces us to either adopt a libertarian theory of moral responsibility, or else give up on the idea of basic desert moral responsibility

[181] As noted in Chapter 1, most skeptical arguments (both anti-compatibilist and anti-libertarian) – including all those mentioned in this chapter – aim to show we cannot satisfy the control condition on moral responsibility, rather than the epistemic condition. So, they question whether we possess the requisite control-conferring (as I have termed them) – as opposed to epistemic – agential features. But some have also challenged whether the epistemic condition can be satisfied. For a useful discussion of this branch of skepticism, and an attempt to rebut it, see Montminy (2021).

altogether.[182] This, skeptics argue, represents the distinctive power of the manipulation argument.[183] Indeed Pereboom and Caruso have themselves presented the manipulation argument as the foundational argument for their skepticism. And as McKenna and Pereboom state, "as regards the dispute between source compatibilists and source incompatibilists, a good deal turns on the force of a manipulation argument like Pereboom's Four Case Manipulation Argument" (2016: 172). Pereboom recently described this as "the best way to argue against the compatibilist" (2021b: 15). So, amongst skeptical arguments, it seems like an especially good candidate for assessing the merits of *SMRT: Retributivism*, *SMRT: Claim Forfeiture* and *SMRT: Basic Reason*. Now to turn to the argument.

4.3. The Manipulation Argument

Manipulation arguments represent a well-established argumentative strategy amongst skeptics. Different versions have been put forward, but each share the same fundamental structure. This structure always begins with a case of outright, intentional manipulation which deterministically causes an agent to act in a certain way whilst satisfying compatibilist conditions. This initial case is designed to elicit a strong intuition of non-freedom and non-responsibility for the agent with respect to their actions. Then, the argument aims to sustain this intuition in a case where the agent deliberates and behaves in the same way, but the deterministic causes are instead those found in a typical deterministic world. I will focus on a particular version of the argument presented by Pereboom, one which I consider especially clear and persuasive. It is also the version most often debated in the literature today.[184]

---

[182] Recognising, of course, that it still remains to be established exactly which version of 'basic desert' the manipulation argument weighs against, if any. Establishing this is the central aim of this chapter.

[183] In particular this explains why it supposedly improves upon the famous Direct Argument, as initially developed by van Inwagen (1983: 188). Cutting a long story short, van Inwagen's argument (like the manipulation argument) purports to show how determinism entails that one's non-responsibility for the past and laws of nature necessarily transfers into non-responsibility for one's actions. But the manipulation argument is purportedly superior because it demonstrates such a transfer of non-responsibility even in a situation where one satisfies *all* compatibilist conditions (McKenna & Pereboom 2016: 154-162).

[184] Manipulation arguments have been presented by, amongst others, Richard Taylor (1974), Robert Kane (1985, 1996), Carl Ginet (1990), Derk Pereboom (1995; 2001; 2014), Noa Latham (2004), and Al Mele (2006, 2008). When I refer to *the manipulation argument*, I should be understood as referring to Pereboom's version, as set out in the text.

Pereboom's manipulation argument centres around an imagined agent named Plum. He is the ideal compatibilist agent – he possesses all those capacities which a compatibilist might imagine to be sufficient for moral responsibility. So, for instance, he is not compelled by any irresistible desires and does not act out of character (Hume 1739; Ayer 1954). He has a harmonious desire structure: he acts on first-order desires which he endorses and identifies with at the higher-order level (Frankfurt 1971). He exhibits a stable and sane pattern of reasons-responsiveness (Fischer and Ravizza 1994; Wolf 1987). He also meets the related condition of being generally able to grasp, apply, and regulate his actions by moral reasons (Wallace 1994). These are just some of the capacities Pereboom has mentioned in his various formulations of the argument. The point however is not to adduce an exhaustive list of compatibilist capacities. Rather, Plum's agential features are an amalgam of *all possible* compatibilist capacities which might be proposed as satisfying the control condition on free will and moral responsibility. The manipulation argument therefore effectively operates with (what I have called) the skeleton concept of agential features, rather than any particularised conception, with the only additional constraint being that they are compatibilist features.[185]

Pereboom's version of the argument is also known as the 'Four-Case' argument, as it asks us to compare our intuitions across four distinct cases or scenarios, sequentially from Case 1 to Case 4.[186] Here are the cases (taken from Pereboom 2014: 76–79):[187]

Case 1: Real-Time Manipulation

> *A team of neuroscientists has the ability to manipulate Plum's neural states at any time by radio-like technology. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know*

[185] Again, the argument does not at all dispute the *existence* of compatibilist agential features, however sophisticated. Rather it questions whether they could endow sourcehood control of the requisite normative significance for moral responsibility. The same is true, it seems, of the majority of skeptical arguments. (The notable exception to this being arguments against agent-causal libertarianism – see Footnote 179).

[186] Other versions of the argument usually draw contrast between only two cases – one of explicit intentional manipulation, and another of 'manipulation' but by typical, non-intentional deterministic causes. Pereboom's use of four cases is distinctive, making the transition between each case more gradual.

[187] Here I summarise Pereboom's argument in my own words, except where indicated (such as in the description of the four cases, where I reproduce Pereboom's own text).

> *will produce in him a neural state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White.*

Case 2: Manipulation at Birth

> *Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic, just like in Case 1. This programming has the intended consequence that in his current circumstances he is causally determined to engage in the egoistic process of deliberation that results in his decision to kill White.*

Case 3: Social Upbringing

> *Plum is an ordinary human being, except that the training practices of his community causally determined his reasoning process so that his deliberative reasoning process is often but not exclusively rationally egoistic (just like in Cases 1 and 2). This training was complete before he developed the ability to prevent or alter these practices. Due to the aspect of Plum's character produced by this training, in his present circumstances, he is causally determined to engage in deliberation that results in him killing White.*

Case 4: Run-of-the-Mill Causal Determinism

> *Everything that happens in our universe is causally determined by virtue of its past states together with the laws of nature. Plum is an ordinary human being, raised in normal circumstances, who is often but not exclusively rationally egoistic (just like in the previous cases). Plum's decision to kill White issues from his strongly egoistic process of deliberation.*

In each case, Plum deliberates in the exact same way in deciding to kill White for egoistic, self-interested reasons. The manipulation that occurs is such that the same, full suite of compatibilist capacities are engaged in his reasoning. He is also assumed to satisfy any epistemic condition on moral responsibility (he knows he is committing serious wrongdoing). The only difference between the cases is the nature of the deterministic causal input which determines how those agential capacities are exercised. These – the nature of Plum's compatibilist agency, and the four cases in which he deploys that agency in killing White – provide the 'raw materials' of the argument.

A question is then posed concerning our intuitions: "is Plum morally responsible in Case 1 in the basic desert sense?" (Pereboom 2021b: 17. See also Caruso & Pereboom 2022: 28). In earlier versions of the argument the question is framed more generally in terms of Plum's moral responsibility, rather than basic desert, such as: "is it plausible that Plum is morally responsible with respect to his decision?" (Pereboom 2014: 75. See also 1995: 24; 2001: 113).

Although Pereboom does not use the term, we are clearly meant to answer this question in light of Plum's *sheer* agential features. Meaning, we cannot draw upon wider normative considerations (for instance, the value which holding him morally responsible or blaming him might serve to cultivate his agency or serve a valuable signalling function).[188] What matters for basic desert is whether Plum's compatibilist agential features do all of the *normative* work, such that they have freestanding reason-giving significance without reliance on wider normative principles.

We are meant to share Pereboom's strong intuition to answer in the negative: Plum *lacks* basic desert responsibility in Case 1. As Caruso and Pereboom say, "whatever sourcehood is at the end of the day, Plum (in Case 1) clearly doesn't have it. And since sourcehood is required for basic desert moral responsibility, we must conclude that Plum is not morally responsible" (2022: 28). This negative assessment effectively comprises the first of the argument's two premises – call this premise 'No Responsibility'. The second premise is that there is "no difference relevant to basic desert moral responsibility between any two adjacent cases" (Pereboom 2021b: 16; Caruso and Pereboom 2022: 26). We are meant to intuit that the difference of deterministic origin seen in each successive case is not the kind of fact which has any bearing on one's basic desert. Call this the 'No Difference' premise.[189]

[188] So, we are testing for an intuition about the existence of a *pro tanto* reason of blameworthiness (grounded in sheer agential features) rather than any intuition about the all-things-considered appropriateness of blame (which would take into account wider reasons).

[189] The wording of the 'No Difference' premise mirrors the concepts we find in the No Responsibility premise. So, in those earlier versions of the argument where the No Responsibility premise was couched simply in terms of moral responsibility (rather than basic desert), the No Difference premise was adjusted accordingly. For instance, it is the denial that there is any "relevant and principled difference

Accepting these two premises entails that our appraisal of Plum's lack of basic desert in Case 1 necessarily transfers through or generalises to each adjacent case. We are therefore forced to conclude that Plum remains equally lacking in basic desert responsibility with respect to his killing of White in Case 4. This reveals how satisfying *all* compatibilist features, as Plum does in Case 4, is insufficient for basic desert. As Caruso and Pereboom conclude: "The manipulation cases serve to indicate that it's possible for an agent who knowingly acts wrongly to be morally non-responsible in the basic desert sense even if the compatibilist conditions are satisfied, and that, as a result, these conditions are insufficient for such moral responsibility, contrary to what the compatibilist claims" (2022: 26).

Pereboom suggests the best explanation for the persisting intuition that Plum lacks basic desert in each case is *simply* "that he is causally determined to act by factors beyond his control" (Pereboom 2021b: 17). Deterministic causation is the common factor in each case. I find this diagnosis very plausible.[190] The argument reveals how determinism of *any* stripe – intentional manipulation or otherwise – prohibits us from being the proper source of our actions. It means that whatever compatibilist agential capacities are engaged, the fact they respond to purely deterministic causal inputs outside of our control means that those capacities *had* to be exercised in the way they in fact were. Plum could not have deployed them in any other way.[191] This denies him the requisite sourcehood control for basic desert.

In my view the power of the argument lies in how Case 1 renders deterministic causal factors especially *salient*. It then sustains this salience as we proceed through to the case of run-of-the-mill determinism, so that we continue to see these factors for the

---

between any two adjacent cases that would show why the agent might be morally responsible in the later example but not in the earlier one" (Pereboom 2014: 75).

[190] Mele (2005, 2006: 141–4) does not, instead arguing that the best explanation for Plum's non-responsibility in Cases 1 and 2 is the intentional, invasive manipulation involved, rather than simply causal determination. This amounts to a form of 'soft-line' reply as discussed below – an evaluation that Plum is non-responsible in Cases 1 and 2, but not in Cases 3 and 4 because the causal determination is of a different, non-excusing quality. Pereboom (2014: 82-86) responds. Whilst I do to not have space to review this, I note that I side with Pereboom here. I will assume then that the best explanation for our negative assessment of Plum's basic desert in Case 1 is indeed causal determination *simpliciter*.

[191] Note this would *not* be the case if we instead posited that Plum possessed agential features which involved indeterminism. In that case it would be possible that Plum might have exercised his capacities differently, and thereby (arguably) gain a greater degree of sourcehood control. Thus the intuition that he is not morally responsible in Case 1 would be less reliably elicited (making space for a hard-line objection). I take it this is why the argument is not taken to have the same force against incompatibilist (libertarian) theories of moral responsibility.

excusing condition they in fact are. (Whereas typically the deterministic causes behind our and others' agency do not possess this salience, and thus we more often fail to appreciate their excusatory force).

Pereboom's argument has of course met with objections. These are usually sorted into soft-line and hard-line replies. Soft-line replies reject the No Difference premise – they posit that there *is* some relevant moral difference between two of the adjacent cases, such that while Plum lacks moral responsibility in Case 1, he might not in (at least one of) the other cases.[192] Hard-line replies instead reject the No Responsibility premise – they grant that the same assessment must apply in all four cases, but insist the correct verdict is that Plum is actually morally responsible in all cases, even Case 1. McKenna (2008, 2014) is the most notable advocate of a hard-line response.[193]

Most debate concerning the manipulation argument has involved offering and responding to such hard and soft-line replies.[194] In my view the argument can withstand both lines of objection. But my aim is not to defend this claim, nor to revise the argument in response to such criticisms. Instead, I challenge the argument from a different direction – a challenge which stands even if we *grant* the argument's conclusion that compatibilist agential features are inadequate for basic desert.

It is important to be clear here about how I take my arguments in this chapter to advance the dialectic with regards to the manipulation argument. There is a substantial literature which debates the effectiveness of the manipulation argument as it has been hitherto presented (most notably Pereboom's version). That debate seeks to establish whether either there exists any successful soft-line or hard-line reply. This literature continues to grow – in a recent series of articles, De Marco and Cyr (2024a; 2024b; 2025) offer a comprehensive summary of the various soft and hard-line replies which have been offered so far. Why do I not engage in more depth with this literature here? Because my aim in this chapter is not to bolster the manipulation argument as it has been hitherto presented against its attackers, whether hard-line or soft-line. As I discuss later (in Section 4.6), those who are already unimpressed with Pereboom's

[192] Other notable soft-line replies include Lycan 1997; Baker 2006; Feltz 2013 and Murray and Lombrozo 2017.

[193] For another hard-line reply see Jeppsson (2020a).

[194] Haji (2009) gives an illuminating summary of hard versus soft-line replies.

argument are not at all likely to be shifted in their position by what I say here. Rather, as will become clear, my chief aim is to identify and to address an unsatisfactory feature of Pereboom's manipulation argument which persists, and poses a problem, *even for those who already impressed by the argument*.

4.4. Conceptual Concerns

My worry about Pereboom's argument concerns conceptual vagueness. Pereboom asks us to evaluate whether Plum is basically deserving, and likewise to affirm that there is no difference between the four cases which could bear upon his basic deservingness. But the concepts invoked in this question are, as we have seen, enormously contested. How are we to understand basic desert here?

Pereboom of course provides *some* account of what he means by basic desert – he gives the updated *Standard Characterisation*. But we have seen how this remains problematically vague both in what it takes blame (and praise) responses to consist in, and in what properties it ascribes to the relevant blameworthiness (and praiseworthiness) reason. Hence Menges', and my own, attempts at clarification. It is only very recently that Pereboom has paired a particular type of reason or worthiness with basic desert, one understood in terms of non-instrumental goodness. But even so, his manipulation argument does not ask whether this (or any other) specific reason is generated by Plum's conduct – it simply asks for our assessment of his basic desert.

The earlier versions of the argument asked us to assess Plum's moral responsibility more generically, rather than specifically in terms of basic desert, but this seems no better. Both philosophers (as we have seen) and the folk alike diverge – or at least commit to a plurality of views – on what exactly it means to be morally responsible.[195] I argued earlier in this thesis that if one is morally responsible for something, this necessarily means that there exists a normative reason to respond – to blame or to

[195] I claim that the folk conceptualise moral responsibility in divergent ways. This certainly seems likely to me, if they even have a stable conception at all (some may have inchoate, unstable notions of what the concept refers to). But this is an empirical claim which would benefit from further experimental testing. Most experimental work in this area has sought to test folk intuitions about the *application* of concepts such as moral responsibility and free will under certain conditions (such as determinism). For instance, see Nahmias et al (2007), Nichols and Knobe (2007), Roskies and Nichols (2008). These studies show a divergence of opinion about when these concepts apply. This might be interpreted as indirect evidence of divergence in how the concepts are being *understood*, but further experimentation is needed here. In any case, it certainly cannot be assumed that respondents to Pereboom's argument have a uniform understanding of what it means to be morally responsible.

praise – at least partly grounded in their agential features and non-neutral moral performance. Here *normative reason* was understood very expansively to include favouring normative considerations of any kind. This includes those considerations which are not necessarily the sort which can figure in rational deliberation, such as the fittingness of (involuntary) reactive emotions. It is eminently possible that respondents to this version of the manipulation argument, in which they assess Plum's *moral responsibility*, might understand the concept in a similarly wide range of senses. (For example, some may be assessing whether reactive emotions seem fitting towards Plum, such as resentment of his quality of regard, or disdain of his quality of character.) All of this means that when a respondent evaluates Plum as being morally responsible (or not) in any of the four cases, this by itself does little to settle which of these various kinds of normative consideration, in favour of which specific response, the respondent has in mind.

If we wish to frame the manipulation argument as a question about Plum's moral responsibility, or basic desert, then we incur a burden to explain what we mean by these concepts.[196] Sommers makes a similar point in relation to arguments in experimental philosophy, both for and against compatibilism, which turn on the concepts of determinism and moral responsibility: "as the philosophical debate is focused on the precise interpretations of these concepts, we ought to devote as much effort as possible to clarifying them for the subjects" (2010: 204).[197] I think this burden has not been satisfactorily discharged in Pereboom's manipulation argument. This means that even if the argument successfully provokes the intuition that Plum lacks basic desert moral responsibility, exactly what this means remains elusive.

Of course, it is inevitably difficult to test folk intuitions about such technical concepts. We could not hope, and so should not aim, for perfect unanimity amongst respondents

---

[196] And notice that in its various presentations, Pereboom's manipulation argument is given without any preamble or further specification regarding how these concepts – basic desert, and moral responsibility – are being understood. The question is simply posed concerning our intuitions about whether these nebulous concepts apply.

[197] Similar concerns about conceptual interpretation have previously been raised more generally about conclusions drawn from experimental philosophy about free will. For instance, Nahmias et al. (2007) argue that the concept of determinism is likely mistakenly interpreted by respondents in Nichols and Knobe's (2007) experimental investigations into free will, and that this undermines the incompatibilist conclusions they claim to derive. My worry about the manipulation argument however is not simply that the relevant concepts (basic desert and moral responsibility) have been misinterpreted by its folk respondents. It is rather that there is not yet even any satisfactorily clear answer as to how those concepts *should* be interpreted according to the argument's proponents.

in how the relevant concepts are being understood. Patterns of intuition elicited by the manipulation argument may well still be useful to establish, even if the concepts in question are being variably interpreted. Indeed, given the conceptual concerns I raise, it might seem curious that many have apparently found Pereboom's manipulation argument so persuasive. It has garnered a pattern of strong negative intuitions about Plum's moral responsibility and basic desert. This result is something which needs reckoning with, and I return to it in Section 4.6.

But given that Pereboom and Caruso seek to motivate skepticism about a very specific conception of moral responsibility, whilst leaving others (such as the answerability and attributability senses) intact, it seems the manipulation argument is not yet properly up to this task. At least, we might aspire to do better. Furthermore, the problem is not merely one of conceptual misinterpretation amongst the folk, but of how the relevant concepts have been under-articulated by the skeptics *themselves*. So, the scope of the argument, in aiming to conclude that *all* compatibilist capacities are insufficient for basic desert, might be impressive. But by drawing upon concepts which are inadequately precise the force of this conclusion is weakened, and its implications left unclear. I think this is a significant flaw in the manipulation argument, and one which has gone underappreciated. I next to turn to see how we might remedy things.

4.5. Testing For *SMRT: Retributivism* and *SMRT: Claim Forfeiture*

Acknowledging these concerns about the manipulation argument, we see how my discussion in Chapter 4 potentially offers a way forward. It established two alternative (more finely-grained) proposals for what basic desert moral responsibility amounts to: *SMRT: Retributivism* and *SMRT: Claim Forfeiture*. So, we might use the raw materials of the manipulation argument to instead test specifically for intuitions about these notions. Can it be used to show that Plum lacks moral responsibility in either of *these* two senses? If so, this promises to serve two purposes. First, to strengthen the manipulation argument in response to the conceptual concerns outlined in the previous section. Second, to vindicate these conceptions of SMRT as capturing what basic desert skepticism (at least its anti-compatibilist strain as exemplified by the manipulation argument) truly gives us reason to be skeptical about.

Leave the raw materials of the argument – Plum's compatibilist agency, and his deployment of this agency in each of the four cases – in place. To test specifically for *SMRT: Retributivism*, we must now ask a different question: *does Plum's conduct in Case 1 give us a reason to blame him in a way which intentionally harms him, because it is non-instrumentally good that we do so?*[198] My intuition here remains steadfastly no – Plum's action and agency, whilst very regrettable, does not render it non-instrumentally good to intentionally harm him with blame. Furthermore, I cannot see why the mere difference in the source of causal determination found in each subsequent case is a fact which could alter this assessment. In other words, the relevant No Responsibility and No Difference premises in this version of the argument both obtain. The conclusion which follows is that Plum's sheer, deployed compatibilist agency seems insufficient for generating this specific kind of reason, even in the natural deterministic scenario of Case 4.[199]

SMRT: Retributivism (unlike *SMRT: Claim Forfeiture*) provides an account of praiseworthiness as well as blameworthiness. On this view, you are praiseworthy where it is non-instrumentally morally valuable to intentionally confer a benefit with praise (on the basis of sheer agential features deployed in positive moral performance). So, we could also run the manipulation argument to test for this specific reason. Imagine that Plum is determined in the same way in all four cases, again exercising his full range of compatibilist capacities. Except this time, he is deterministically caused to deliberate unselfishly (rather than egoistically) and to courageously rescue White from peril. Once again, my view is that in Case 1 there is no sense in which

[198] Plugging *SMRT: Retributivism* into the manipulation argument in this way would seem a natural next step for skeptics such as Pereboom and Caruso. That is, if they wish to maintain a) that it is specifically intentional harms and benefits (accountability responses) inflicted for reasons of non-instrumental goodness (which is how Pereboom at least latterly understands basic desert) which is the target of their skepticism. And b) that the manipulation argument is their primary skeptical argument. It is curious therefore they have not done so, but I am happy to take this step on their behalf.

[199] My sense is that the intuition at play is actually more general than this. It is that we lack *any* normative reason to intentionally harm Plum sheerly in virtue of his compatibilist agency, rather than merely and more specifically a reason of non-instrumental goodness. So, this more general intuition about absent reasons naturally includes reasons expressed in terms of non-instrumental goodness. But it concerns *all* prospective normative reasons, without discrimination. (Including for instance those which might be expressed in deontic rather than axiological terms, such as that it would be right or dutiful – rather than good – to intentionally harm Plum. For example, Feinberg describes reasons of desert in such 'ought' terms: "[t]hat a subject deserves X entails that he ought to get X in the *pro tanto* sense of 'ought'" (1970: 60). Or consider Kant's famous view that murder generates a strict moral duty to capitally punish the offender, even if society is dissolving (1797: 102)). Noticing this *generality* in the intuitions provoked by the manipulation argument is a central theme in the arguments of this chapter, as will become clearer.

Plum's morally good action can explain why it is non-instrumentally good to intentionally benefit Plum via praise.[200] Such benefits might take the form of rewarding Plum with public approbation or financial reward, say. Intuitively, Plum in Case 1 is not the source of his actions in any way which could make sense of his having deserved or *earned* such benefits. Furthermore, the differential deterministic input in each case is not something which alters this fact about Plum, and so this verdict persists through to Case 4.

Therefore, I submit that the manipulation argument is effectual against these praising responses *too*. In fact, Pereboom (2014: 102) briefly discusses how he thinks his original argument "works equally well" against basically deserved praise. He is responding to Nelkin (2011: 56), who suggests it "is much less appealing to say that Plum is not responsible" in a case where Plum responds to the right reasons and acts well, even when deliberately manipulated to do so. We can certainly agree with this. There are plausibly many reasons to respond positively to Plum's laudable conduct – to consider it *praiseworthy* – and it would be a miserable world where good deeds (even those which are the product of deterministic forces) went entirely unpraised and unrewarded. But this revised manipulation argument indicates how none of these are forms of praiseworthiness grounded in non-instrumentally good benefits, and so Plum lacks the specific mode of praiseworthiness at issue.[201]

We therefore have grounds to think that the full range of reasons and responses implicated in *SMRT: Retributivism* lack any application to Plum. His compatibilist agential features entirely lack the kind of normative significance characterised by this view.

How does Menges' view – *SMRT: Claim Forfeiture* – fare in light of the manipulation argument? We can test for the response-type which characterises this view, with a suitably modified question: *has Plum in Case 1 forfeited any standing claims against a non-trivially harmful response and do we have a positive reason to respond in this*

[200] Note, it might still be non-instrumentally good to benefit Plum – presumably, to confer a benefit on anyone, thus making their life go better, is non-instrumentally valuable for that person. The point is that no *new* reason to benefit Plum in *particular* is introduced by his conduct – it does not make benefitting him any more non-instrumentally valuable than it already was.

[201] One might well argue that we do not need to provide any reason for these praising responses – unlike intentional harms, they are not *prima facie* wrong. Maybe so, but that is orthogonal to what is at issue here, which is simply whether there is such reason.

way? Once again, I am inclined to answer negatively – if Plum is intentionally manipulated to commit wrongdoing, as in Case 1, then surely it is false that he thereby forfeits any claim against being non-trivially harmed. That would seem to pile injury upon his misfortune. Again, there is no plausible difference in Cases 2 - 4 which could affect the status of Plum's claims against non-trivial harm. The conclusion is that Plum is not basically deserving in Menges' sense, even in the naturally deterministic circumstances of Case 4.[202]

Note, this negative verdict applies to *both* components of blameworthiness described in *SMRT: Claim Forfeiture*. Plum's conduct does not furnish any positive reason to inflict non-trivially harmful treatment on him; *nor* does it even mean he forfeits any claim against such maltreatment. His conduct does not shift the balance of reasons in either of these ways. So, compatibilist agential features lack even the more modest eliminative power to cancel existing reasons against inflicting non-trivial harm, never mind the power to generate new positive reasons to do so. This finding bears particularly, I think, upon two distinctive theories of desert recently offered by Dana Nelkin and Doug Portmore. Nelkin (2016) argues that desert constitutes a conditional reason to inflict harm, one which only becomes effectual under certain conditions. She mentions only one condition:

> Suppose that you are in a position in which you have no choice but to inflict harm and you can harm someone deserving of it or harm someone else. This may give you a reason to harm the person who is deserving rather than others (Nelkin 2016: 179).

We can translate this into the language of claim forfeiture. For Nelkin, your deservingness does not by itself constitute a positive reason to intentionally harm you, but it does entail you forfeit a certain moral claim. Namely, one which says your claim against being intentionally harmed is equally as strong in all situations as that of others. Portmore (2024) presents what I take to be a kindred view. He thinks that the

[202] Menges (2023: 631) himself suggests reframing the manipulation argument in these terms. But he has independently decided (using the desiderata I criticised in Chapter 3) that *SMRT: Claim Forfeiture* is the correct conceptualisation of basic desert. He recommends that we then test whether the manipulation argument weighs against this conception of basic desert, since he thinks this is what any *bona fide* basic desert skeptic *should* be seeking to disprove. I argued in Section 3.3.5 that this gets things the wrong way round. The point is to use the manipulation argument to *determine* how to understand the phenomenon in question (the kind of blame skeptics should be skeptical about), whether this is best understood as a version of basic desert or something else.

blameworthy deserve to suffer pain; it is non-instrumentally good that they do. He argues that the probative comparison for establishing this is not whether the blameworthy should suffer or not, but whether they or the *non*-blameworthy should suffer. He suggests it is much harder to resist the intuition that the blameworthy should suffer rather than the non-blameworthy.[203] Put in terms of claim forfeiture, Portmore is defending a form of blameworthiness wherein one's claim against suffering is weaker compared with others (the non-blameworthy).

Each of these views therefore involve forfeiting claims to equal protection against non-trivial harm. But my assessment of Plum, in Case 1 and then sustained through to Case 4, is that he does *not* forfeit such claims. He does not (in virtue of his sheer compatibilist agency) forfeit his equal claim against harmful treatment, even in a situation where harm must be inflicted on someone. Nor does he have a weaker claim against suffering in comparison to anyone else, even the non-blameworthy.[204] (One line of argument, which I will mention but not pursue, might even suggest the opposite is true. Plum has already arguably suffered great, harmful misfortune in having his agency coerced – by deterministic forces – into serious wrongdoing. Perhaps considerations of fairness mean his claim against (further) harm now merits *special* protection in comparison with those others whom determinism has treated more generously, coercing them into less morally heinous behaviour).

So, whilst I argued in Chapter 3 that the desiderata used by Menges are inadequate for vindicating his conception of basic desert, the manipulation argument *does* now appear to provide this vindication. Its raw materials can be martialled to cast doubt upon the theory of moral responsibility described in *SMRT: Claim Forfeiture*, as well as that described in *SMRT: Retributivism*. Are *both* views therefore vindicated as accurately

[203] Portmore thinks our intuitions about this comparison are more reliable because both options involve the same quantum of suffering; it is just a question of its distribution. Whereas asking whether the blameworthy should suffer or not asks us to choose between two worlds with unequal amounts suffering. Portmore argues that this confounds things. Even if we had the intuition that it is non-instrumentally good (because more just) that the blameworthy suffer, we might simultaneously think this goodness is outweighed by the badness of there being more suffering in the world (since all suffering is still non-instrumentally bad). In which case we might prefer the world without any suffering, despite it being a world which contains less justice. As he summarises, "[t]he point is that something can be good in some respects but bad in others. And if the bad outweighs the good, this something will be overall bad" (2024: 802).

[204] I have misgivings about Portmore's proposed comparison between the blameworthy and non-blameworthy, which I return to in Section 4.7.2.

capturing what it is that the skeptical considerations of the manipulation argument undermine?

A quick reply: yes, but the vindication they receive remains somewhat unsatisfactory, given the unsettledness inherent to both accounts, as discussed in the previous chapter. Even if we establish that sheer compatibilist agential features cannot ensure that intentionally harmful blame and beneficial praise is non-instrumentally good, we are still left with some conceptual ambiguity in how we are to understand blame and praise. And, whilst less conceptually ambiguous, even if we establish that sheer compatibilist agential features cannot generate a reason to inflict non-trivial (claim-infringing) harm, this is a category of responses with in-built conditionality. It is conditional upon which responses our moral theory deems to be *non-trivially* harmful, and within which contexts. (Plus, it is unclear here what we should say about praiseworthiness). In different ways, then, this vindication of both *SMRT: Retributivism* and *SMRT: Claim Forfeiture* still leaves it somewhat inconclusive as to precisely which responses are being undermined.

But I think we can say more. I will now make a case for *SMRT: Basic Reason* in fact being the superior articulation of SMRT – whilst explaining why it is these other two SMRT proposals appear also to be vindicated by the manipulation argument.

4.6. Asking A Different Question – The Pertinence of *SMRT: Basic Reason*

Here is a new question we might ask about Plum: *In Case 1, do Plum's actions and his agency provide a normative reason of any description to respond with an action of any kind?* This is the question which *SMRT: Basic Reason* invites us to ask. Once again, we are to take only Plum's actions and agency, and its deterministic origin, into account, and no further normative considerations. Can these facts do *all* of the normative work in generating a reason for action, or are additional normative principles essential?

My strong intuition here, once again, is *no*. Reflecting upon Plum's actions in Case 1, my sense is not merely that we lack reason to intentionally harm Plum, or to infringe any of his other moral claims. Rather, his exercise of agency, given its origination in the neuroscientists' real-time manipulation, is *entirely* without freestanding reason-giving significance. By which I mean, it does not generate any normative reason recommending an action which would be appropriate in response. There is no blameworthiness of this sort.[205]

Since we are testing for *SMRT: Basic Reason*, the argument's No Difference premise is also broadened accordingly. Is there any difference between adjacent cases which could alter the balance of all normative reasons for action? Here too I cannot see that the different variety of deterministic causation found in each case by itself introduces any new normative considerations. Thus, in virtue of the same generalisation strategy used in Pereboom's original argument, I evaluate Cases 2, 3 and 4 in an identical way. Even where it is a product of run-of-the-mill deterministic forces, Plum's action lacks any freestanding reason-giving significance. (My intuitions are the same if we run the argument again but instead imagine that Plum acts virtuously, rescuing White in each case, as we did for *SMRT: Retributivism*. Plum's compatibilist agency, given its origin in deterministic forces, fails to generate any praiseworthiness in the sense of providing any normative reason to act one way or another). Furthermore, I do not see how Plum's action in Case 1 serves *by itself* – without reference to any wider normative principles – to weaken or eliminate any *existing* reasons we might have to act in any particular way**.**

So, I claim that this iteration of the manipulation argument tells against *all* putative normative reasons for action grounded in sheer compatibilist agential features. Plum satisfies *all* possible compatibilist criteria – yet this is inadequate for generating any consideration in favour of any action. This therefore reveals the pertinence of *SMRT: Basic Reason* – it accurately describes what it is which these skeptical considerations (the raw materials of the manipulation argument) give us cause to be skeptical about. Seen this way, the manipulation argument delivers a more powerful dose of skepticism than has previously been claimed, in that its scope encompasses *all* reasons for action and not merely the more specific kinds of reason highlighted by *SMRT: Retributivism* and *SMRT: Claim Forfeiture*.

---

[205] To be clear, here I use *blameworthiness* in my stipulated sense from Chapter 1. This understands it as a status which provides a normative reason to respond with blame, which is at minimum grounded in agential features deployed in subpar moral performance.

I take this to imply that facts about our compatibilist agency are, at least in one important sense, no more normatively significant than facts about our height, for example. We would much more readily disregard Plum's height as a feature which could, in combination with his wrongdoing, generate a reason in favour of any particular action. It is hard to imagine what sort of reason this would be, absent any additional normative considerations. It remains true that compatibilist agential features are in another sense more normatively potent than facts about our height. They can be much more readily *combined* with additional considerations in order to generate certain kinds of normative reason. This explains why those very agential features – most notably, reasons-responsiveness – tend to figure in instrumentalist theories of moral responsibility.[206] For instance, the fact that Plum is reasons-responsive could explain why it would be *instrumentally* advantageous to respond to his action in certain ways – perhaps even to infringe his standing moral claims, for instance by administering painful punishment. Such responses might influence Plum to exercise his reasons-responsiveness, and the control this affords him over what he does, towards more virtuous ends in the future. But because our height is not a feature which confers us agential control of any kind over what we do, it is much less clear how we might exploit this feature of Plum in our responses to him, with a view to affecting his future decisions.[207] So, compatibilist agential features remain hugely, and distinctively, important for *influencing* the reasons we have for responding to each other. But (like height) they are entirely unsignificant in the sense which is relevant here – the capacity to generate reasons independently.

The claim that compatibilist agential features entirely lack freestanding reason-giving significance of this kind may surprise. But it is worth being explicit about what I am *not* claiming. As discussed in Chapter 3, *SMRT: Basic Reason* strictly concerns normative reasons for *action* (or omission) – the kinds of consideration which might figure in our deliberations about what to do. I do not dispute that Plum's sheer compatibilist agency may yet retain normative significance of a different kind to this,

---

[206] For instance, see Jefferson (2019), McGeer (2015), McGeer and Pettit (2015), and Vargas (2013) for (indirect) instrumentalist accounts which are based upon reasons-responsiveness.

[207] Still, one's height is a fact which *can* – in combination with normative principles – contribute to normative reasons of certain other kinds. For example, if we aim to recruit for a successful basketball team, then the fact that Scottie is 7ft tall is an important part of the normative reason which favours selecting him for the team.

concerning other kinds of praise and blame response.[208] For instance, it can clearly still generate reasons to form *beliefs* (such as, in all four cases, the belief that Plum has murdered White). Also, reactive *emotions* towards Plum may yet be fitting. For instance, perhaps resentment or disdain of Plum is fitting in some or all of the cases, simply in virtue of his actions and agency).[209] Neither of these represent the kind of normative consideration in question – reasons in favour of action – which is what I exclusively claim is undermined by the manipulation argument.

Also, in practice myriad normative reasons to act will of course be generated once we learn of Plum's conduct, in all four cases. For instance, we will have good reasons to steer well clear of him, to alert others to his murderous conduct, perhaps to incarcerate him and possibly even to violate his claims against non-trivial harm in other ways too. The different deterministic story present in each case – whether Plum was determined to murder White by malicious actors, or instead by naturalistic deterministic forces – *will* impinge on the configuration of *these* reasons. (For instance, these facts will inform the kind of modifications we ought to make to his environment, if we wish to inhibit him acting murderously again). But my central claim is that none of these

[208] Here I use praise and blame according to my very expansive conception of these terms, as set out in Chapter 2. This places no limit the kind of response which might qualify as blame and praise (so, including emotions and even beliefs). The only condition is that they are a response to the appraisal of moral over- or underperformance resulting from deployed agential features.

[209] Whilst I do not take any firm stand here, my view is that compatibilist agential features lack normative significance in one of these respects, too. They are insufficient to make fitting the reactive emotion of moral anger. This runs contrary to the hugely influential view of moral anger spearheaded by PF Strawson. The Strawsonian view says that what makes moral anger fitting is a singular compatibilist feature: poor quality of will. In particular, where one shows "ill will or indifferent disregard" towards another (Strawson 1963: 200. See also Shoemaker (2015: 87-117) for a fuller discussion). So, this view deems moral anger towards Plum to be fitting (certainly in Case 4), given he undoubtedly exhibits poor quality of regard towards White, even whilst he is entirely causally determined to do so. I think this view misses out something essential about anger's fittingness – and that the manipulation argument might be martialled to reveal this, too. In the case of intentional manipulation in Case 1, I submit that anger towards Plum feels entirely inappropriate. It will naturally subside once we learn that his actions were intentionally manipulated to occur. This diminishment of anger is sustained if we then work successively through the cases to Case 4 in the usual way. The salience of deterministic causes serves in each case to undermine the tendency to feel anger. What I take this this to imply is that whilst poor quality of regard is plausibly *part* of what accounts for anger's fittingness, this is not the whole story. It also requires something else: an *ignorance* of any deterministic causes which produced that lack of regard. To feel anger, we must suppress of our awareness of such causes. But the manipulation argument makes us confront these causes with regards to Plum, and thereby reveals how anger's fittingness depends upon our ignorance of them. It would require a separate argument to properly substantiate these claims, which for reasons of space I do not pursue (but aim to take up in separate work). For now, I sketch them to indicate how the manipulation argument might be mobilised to undermine an even wider range of responses, revealing how compatibilist agential features may lack further kinds of normative significance beyond those which are my central focus.

normative reasons will be *basic* in the relevant sense. None derive sheerly from the fact that Plum's agential capacities have been exercised in wrongdoing owing to deterministic causation. Instead, they will each draw upon additional normative considerations (further moral reasons or principles, or our prudential interests or legal norms, say).

Still, the claim that compatibilist features lack *all* freestanding reason-giving significance is bold. One might resist it by again adopting either a hard or soft-line position (the former insisting that Plum's agency and action *does* generate some reason for action in all four cases, the latter insisting that it does in at least Case 4). Also, those who disagree with the arguments of the previous section – that *SMRT: Retributivism* and *SMRT: Claim Forfeiture* are vindicated by the manipulation argument – will *necessarily* object now. They already think there exists a reason to respond to Plum with a blaming action of at least one sort (either intentional harm or else some other non-trivially harmful response).

I do not offer any decisive refutation of such replies. Much here turns on intuitions, and further investigation is required. Rather, I foremostly offer a provocation. I present three revised versions of the manipulation argument, inviting us to test our intuitions regarding three distinct categories of response as expressed by *SMRT: Retributivism*, *SMRT: Claim Forfeiture* and *SMRT: Basic Reason*. Each of these asks us to assess the application (to Plum) of concepts which are significantly less ambiguous, to philosophers and folk alike, in comparison with Pereboom's original manipulation argument which (I argue) is beset with conceptual difficulties. Indeed *SMRT: Basic Reason* brings hardly any conceptual complexity at all – it enquires about the existence of *any* reason for *any* action, without differentiation. If there is less scope for conceptual misunderstanding here, then folk responses to this version of the manipulation argument might be especially useful to garner.[210]

Whether or not one shares my intuitions, then, I hope at least to have shown which intuitions we might more profitably investigate. Reframing the manipulation argument

[210] Although of course it still requires that respondents satisfactorily comprehend those concepts which make up the 'raw materials' of the manipulation argument – the nature of deterministic causation, and Plum's compatibilist-only agential capacities.

in these ways promises to deliver greater clarity about which specific reasons, in favour of which actions, sheer compatibilist agential features can generate (if any).

Still, I do think I can somewhat fortify my claim that the manipulation argument is effective against all reasons for action (as per *SMRT: Basic Reason*) by way of responding to the following possible objection. It might be pointed out that Pereboom's original formulation of the manipulation argument undoubtedly possesses power, despite the conceptual concerns I raise. It is so well-established in the literature because many *already* find it somewhat persuasive in telling against compatibilist moral responsibility and/or basic desert. Even dissenting authors have at least been willing to respond to Pereboom in like conceptual terms: hard and soft-line replies seek to defend Plum's status (in Case 4 at least) as morally responsible or as basically deserving. This might suggest that there is greater convergence about what these concepts mean than I have claimed. Perhaps I have overstated the conceptual worries and the need to reframe the manipulation argument in response to them?

In reply I suggest the vindication of *SMRT: Basic Reason*, rather than being rendered irrelevant or superfluous by the power of Pereboom's original argument, provides a ready *explanation* of that power. As discussed, it is very likely that respondents who agree with Pereboom that Plum lacks moral responsibility or basic desert are operating with divergent understandings of these concepts. If the manipulation argument is effective in cancelling *all* putative normative reasons for action on the basis of sheer compatibilist agential features, then this divergence will not matter. Whatever kind of normative reason these respondents have in mind, to go in for whatever kind of response, that reason will be cancelled. So, we have a hypothesis for how it is that Pereboom's version of the manipulation argument can successfully provoke shared assessments of Plum's non-responsibility and non-basic desert, despite the ambiguity inherent to these concepts.[211]

Similarly, this would also explain how we can be confident that the same skeptical considerations weigh against intentionally harmful blame and beneficial praise as

[211] This hypothesis assumes that these respondents to Pereboom's original argument are at least understanding moral responsibility and basic desert in terms of *reasons for action*. It is possible some might have other kinds of normative significance in mind (the fittingness of reactive emotions, for example). It will be important to identify and disentangle these different conceptual interpretations, again by explicitly testing for intuitions about those specific response-types.

conceptualised in *SMRT: Retributivism*, despite some unclarity about which precise actions it takes to qualify as blame and praise responses. Likewise it explains how we can be confident that we lack any positive reason *whatsoever* to inflict non-trivial, claim violating harm (as per *SMRT: Claim Forfeiture*) on Plum. The power of the manipulation argument in both cases trades upon its more general power to discredit all reasons for action, inclusive even of yet-to-be-specified reasons in favour yet-to-be-specified actions.

Suppose we grant that the manipulation argument is effective in undermining each species of blameworthiness expressed by *SMRT: Retributivism*, *SMRT: Claim Forfeiture* and *SMRT: Basic Reason*. I propose we need not reach for any new explanation for these intuitions. The best explanation remains that it is causal determination *simpliciter* which inhibits these respective forms of blameworthiness in Plum, across all four deterministic scenarios. Again, the simple power of the manipulation argument lies in rendering deterministic causal factors explicitly salient in Case 1. This salience impresses on us how deterministic circumstances have *exploited* Plum's agency for ill, and this serves immediately to undercut the normative significance of Plum's agential sophistication in opting to murder White. His agential features – his reasons-responsiveness and quality of will, for instance – certainly confer him some measure of control over his actions.[212] But the fact these features are deployed according to the neuroscientists' malicious intentions seems clearly to deprive him of the *requisite* sourcehood control for all three forms of blameworthiness. Then, in each subsequent case, the nature of the deterministic causal origin (whether intentional or not) is something which Plum equally and entirely lacks control over. There is no room therefore for the necessary sourcehood control to enter at any point.

I therefore argue that both Pereboom and Caruso's *SMRT: Retributivism*, and Menges' *SMRT: Claim Forfeiture*, correctly identify kinds of moral responsibility which the skeptical considerations of the manipulation argument undermine. The same goes for *SMRT: Basic Reason*. Are these three views therefore on a level footing, or is there reason to think that one represents the *optimal* way to describe what it is that the manipulation argument debunks? In the final section I offer some thoughts on this, as

[212] This is the point I emphasised in Section 1.3.4 of Chapter 1 – the skeptics do not query that these agential features exist, only whether the control they confer is adequate for basic desert moral responsibility.

well as the wider implications of my arguments for how we should (re)conceive basic desert skepticism.

4.7. Reconstruing Moral Responsibility Skepticism

I finish by considering two broad sets of implications of this preceding discussion for how we should think about basic desert skepticism. First, those which follow if we accept my argument that *SMRT: Basic Reason* is vindicated by the manipulation argument. Second, implications which follow even if we do not, concerning the methodology I have employed. I will take these in turn.

4.7.1. Implications of the Vindication of *SMRT: Basic Reason*

If the manipulation argument does indeed vindicate *SMRT: Basic Reason*, this implies the object of the skepticism pushed by the manipulation argument is *recentred*. It is no longer best understood as skepticism about the propriety of any specific kind of blaming action (intentionally harmful blame, say). Nor as concerning the existence of any specific kind of blameworthiness reason (one expressed in terms of desert, non-instrumental goodness, or otherwise). In fact, we need not even invoke the concepts of blame or blameworthiness. Rather we can express this skepticism using the sparser conceptual apparatus of *SMRT: Basic Reason*, which simply refers to normative reasons for action.

The object of this skepticism is now seen as the *normative potential of sheer compatibilist agential features*. Its central claim is that these features, when deployed in non-neutral moral performance, do not by themselves generate *any* normative reasons in favour of any action. Skepticism about those more specified actions, and reasons for action, still stands – but it is now seen as a downstream implication of this more fundamental skepticism. The position is more aptly labelled as skepticism about *basic reasons*, rather than basic desert, since this foregrounds how sheer compatibilist agential features are unable to generate any reasons for action whatsoever, desert-based or otherwise.

This is not merely a question of the optimal way to describe the skepticism in question, however. More consequentially, as I indicated earlier in Section 3.4, it means that skeptics are liberated from having to fight some difficult conceptual battles. Were it any less than all normative reasons for action which were undermined by the manipulation argument, then we may well need to use more fine-grained concepts to describe the relevant set of responses. But since all normative reasons are excluded, we need not. So, for instance, skeptics need not be exercised over what distinguishes blame and praise from other kinds of response. Nor do they need to give an account of what reasons of desert entail (or the attendant question of which responses can qualify as 'deserved'). Nor even to commit to any particular conception of moral responsibility. Again, their skepticism can be more sparsely put in terms of (the total absence of) normative reasons for action.[213] However, once we present the object of skepticism as a particular species of blame and praise (as *SMRT: Retributivism* does), or as a type of desert or moral responsibility, then we incur a burden to adequately explain what we mean by these concepts. I argued previously that this a burden which the basic desert skeptics Pereboom and Caruso have not met. Happily for them, we now see it is not a burden that they need to carry. It enters them into conceptual complexities which they can avoid, and it obscures the manipulation argument's full scope. Instead, it is the burden of the skeptics' *opponent* to demonstrate how sheer agential features can generate specific kinds of reason in favour of a specific, and proportionate, response (as McKenna, to his credit, attempts to do).

Where does this leave *SMRT: Retributivism* and *SMRT: Claim Forfeiture*? I argued that both views are also vindicated by the manipulation argument, in that the argument serves to undermine the categories of response which they identify. However, it does so only in virtue of undermining the more general class of *all* normative reasons in favour of action, in line with *SMRT: Basic Reason*. What I think this shows is that *SMRT: Retributivism* and *SMRT: Claim Forfeiture* are responsive not merely to the skeptical considerations of the manipulation argument. They also respond to something else: our *moral concerns*. The kinds of response which these views respectively highlight – intentionally harmful blame and beneficial praise; and blame which (more generally) entails non-trivial, claim-violating harm – are *prima facie*

[213] In doing so we likewise avoid the inconclusiveness which, as discussed, accompanies *SMRT: Claim Forfeiture*. On that view, in order to decisively establish the set of responses which skepticism discredits, we need to identify which of our actions amount to non-trivial, claim-violating harm. (This in turn may require that we attend to contextual variability of the sort described in Section 3.3.1). Whereas, if *all* normative reasons for action are discounted as per *SMRT: Basic Reason*, the scope of the skepticism is more immediately and conclusively settled.

morally problematic.[214] They are the kinds of response which we might worry are in poor moral standing. It is therefore particularly *important* that we establish whether sheer agential features can generate a reason in their favour.

This is why I think *SMRT: Basic Reason* is the superior SMRT proposal, most successful in identifying *the kind of blame which skeptics should be skeptical about* (to borrow Menges' phrase). We have Menges' proposal: any response (to sheer agential features) which would otherwise violate a standing claim we have against such treatment. But I take Menges to be giving an answer – a plausible one – to a subtly different and somewhat narrower question. Namely a question about the kind of blame which skeptics should be skeptical about *and should also especially care about*.[215] This is no longer purely a matter of establishing the conclusion which the relevant skeptical considerations deliver – it is also a matter of which part of this conclusion especially matters to us. These are two categorically different considerations, and we should disaggregate them. It is *SMRT: Basic Reason* which corresponds purely to what the skeptical considerations of the manipulation argument establish.

---

[214] Harmful responses are of course typically seen as more morally problematic – in the sense of demanding moral justification – than beneficial ones. This is likely one important reason why blame (typically associated with harming) has received more attention in the literature than praise (which is usually taken to be a beneficial, or at least benign, kind of response). Susan Wolf, for example, expresses this line of thought:

> I think we have stronger reasons for wanting acts of blame to be justified. If we blame someone or punish him, we are likely to be causing him some pain. But if we praise someone or reward him, we will probably only add to his pleasures. To blame someone undeservedly is, in any case, to do him an injustice. Whereas to praise someone undeservedly is apt to be just a harmless mistake (1980: 156).

(For further discussion on the lack of attention which praise has received relative to blame, and why this might be, see Holroyd 2021: 2; Lippert-Rasmussen 2022: 230 n.4; Telech 2022: 1–2). But even beneficial praise might raise moral concerns – perhaps most obviously with respect to the justness of its distribution. Holroyd has notably critiqued *oppressive* praise: that which is "systematically, unjustly apportioned […] in ways that reflect and entrench existing structures of oppression" (2021: 1). Other recent work has drawn attention to further harms which praise might cause, both to its recipient and to the wider social group to which they belong, even where it is intended as a benefit (see for instance Bingeman 2024). So, whilst intentionally beneficial praise is perhaps less morally concerning than intentionally harmful blame, it is not necessarily harmless. We therefore have a moral interest here too in establishing whether sheer compatibilist agential features can generate reasons in favour of such responses (or not, as Pereboom and Caruso, and I, contend).

[215] Perhaps, being charitable, it is this narrower question which Menges intended to answer all along. (In which case his revisionary and justice constraints might likewise reflect a narrower focus on responses which should morally concern us). This might be a reasonable enough intention – but it is not suggested by Menges' stated aim of simply finding "the kind of blame skeptics should be skeptical about" (2022: 1).

The narrower categories of response identified by *SMRT: Retributivism* and *SMRT: Claim Forfeiture* therefore reflect something important. But this importance is now seen in more practical terms. They highlight the *egregiousness* of the error we risk making in our practices. Not only do we mistakenly imbue compatibilist agential features with freestanding reason-giving significance – we imbue them with such exaggerated significance that we take them to generate reasons in favour of such morally questionable conduct.[216]

How widespread is this egregious error? Menges is surely correct that our interpersonal practices *do* frequently credit sheer agential features (unlike height) with the kind of normative significance described in *SMRT: Claim Forfeiture*. That is, our practices presuppose that they *can* generate reasons to respond in non-trivially harmful ways. He gives the examples of aggressively confronting another for breaking a promise, or harshly reproaching someone for betraying their friend (Menges 2022: 12). Such commonplace responses are plausibly ones which we have standing moral claims against being subjected to – and also ones often taken to be justified *simply* because of wrongdoer's deployed agential features (such as their poor quality of will or failure to adequately respond to reasons). Furthermore, the kinds of response identified in *SMRT: Retributivism* undoubtedly figure, to some extent, in both our interpersonal and institutional practices. As discussed in Section 3.3.4, I am much less confident than Menges that our interpersonal responses do not sometimes involve the retributivist intention to inflict harm. Certainly, at the institutional level, legal punishment is often (at least partly) motivated by retributivism. The United States provides one clear example: the American Law Institute recently revised the Model Penal Code, reinforcing desert as the official dominant principle for sentencing.[217] The

---

[216] In this respect, I suggest *SMRT: Claim Forfeiture* is the preferable account to *SMRT: Retributivism*. If it is a matter of picking out those responses which are both in scope of the manipulation argument *and* which are also of especial moral concern, then we ought to be interested in the full set of those responses which threaten to infringe our interpersonal moral claims, without limit. This includes non-intentionally harmful blame (which *SMRT: Claim Forfeiture* accounts for) as well as the intentionally harmful blame on which *SMRT: Retributivism* exclusively focuses. Similarly, consider *SMRT: Claim Forfeiture's* non-specificity with regards to the nature of the positive reason to blame which it takes to be entailed by basic desert. This non-specificity is preferable to *SMRT: Retributivism*'s specification of a particular reason (one understood, at least by Pereboom, in terms of a particular kind of value: non-instrumental goodness). We ought to be interested in the full set of reasons which the manipulation argument weighs against, and this (so I claim) amounts to *all* putative reasons for action, whether grounded in non-instrumental goodness or otherwise.

[217] American Law Institute, Model Penal Code Sect. 1.02(2) adopted May 24, 2017.

Supreme Court also affirms retributivism as the "primary justification for the death penalty" (Robinson 2008: 145–146). Equally, consider rewards which are doled out for moral achievements such as the Nobel Peace Prize or George Cross (two particularly esteemed examples). These are typically justified in terms of their recipients' desert, given how they have exercised their agency. The vindication of *SMRT: Retributivism* and *SMRT: Claim Forfeiture*, then, promises to have substantial revisionary implications for all such practices.[218] It discredits one important kind of reason which is used to justify them.

A more precise assessment of these practical implications must be left to future work.[219] But my analysis of the manipulation argument recommends a singular test for the initial interrogation of our practices: where do they involve the normative error of taking deployed compatibilist agential features to generate, by itself, a reason to act? The *ensuing* task is then to consider which of these practices we ought to attend to, and potentially revise, most urgently – and this is where *SMRT: Retributivism* and *SMRT: Claim Forfeiture* become valuable. They highlight which of these error-involving responses are at particular risk of being morally unjustified.[220]

[218] Indeed as discussed in the previous chapter, when it comes to the practical implications of their skepticism, we see that the basic desert skeptics have focused their efforts on proposed revisions to criminal justice.

[219] Also note that, if we accept that compatibilist agential features cannot by themselves generate any normative reason for any action, this does not *necessitate* that we should revise or abandon those practices which are justified using those reasons. Perhaps those practices can be justified on other grounds (such as their beneficial consequences or using contractualist ethical resources). Alternatively consider Nichols, Timmons and Lopez's (2014) notion of "ethical conservatism" and Smilansky's (2000) recommendation of "illusionism". These are distinct views, but both share the idea that we would be better off preserving our responsibility practices even if they are premised on bad reasons, given how entrenched they are, the valuable role they now play and how damaging it would be to leave them behind.

[220] Since these highlighted responses – certainly the harmful ones – involve higher moral stakes, our belief in the existence of reasons in their favour plausibly ought to meet a higher epistemic standard. As Benjamin Vilhauer persuasively argues, "[i]f it can be reasonably doubted that someone had free will with respect to some action, then it is a requirement of justice to refrain from doing serious retributive harm to him in response to that action" (2009: 131). We might expand this principle beyond serious retributive harm to include more modest, yet still non-trivially harmful responses. Here too we need to be *especially* confident in our belief in reasons favouring such responses before we use those reasons to justify our practices. I have argued that the manipulation argument should fully erode our confidence in all reasons based on sheer agential features in favour of *any* kind of action. But when it comes to reasons favouring those harmful kinds of action identified by *SMRT: Retributivism* and *SMRT: Claim Forfeiture*, even a slight erosion in our confidence should weigh heavily in our deliberations about what to do. Similar arguments for greater epistemic humility when it comes to harmful responses have been made by Pereboom (2001), Sofia Jeppsson (2020b) and Elizabeth Shaw (2021).

4.7.2. Implications of Methodology: Moving Away from Blame and Blameworthiness

I have so far considered the implications which follow if we accept the core claim of this chapter: that the manipulation argument weighs against *all* normative reasons grounded in sheer compatibilist agential features. But this admittedly remains a provisional claim – the relevant intuitions may well differ, and further investigation into these intuitions is needed. However, even those who are unconvinced by this claim might take something from the methodology I have used in reaching it.

This methodology is straightforwardly summarised: it involves jettisoning concepts which are ambiguous and under-defined, if these can be replaced by sharper ones which more precisely capture the relevant considerations. When it comes to the manipulation argument, this means dispensing with concepts such as moral responsibility, blame, blameworthiness, and even desert.[221] We can instead reframe the argument as one about the relationship between agential features and different categories of normative reasons for action. By doing so, whether we agree with the conclusion of this reframed argument or not, we render much more clearly the particular kind of skepticism at stake.

I believe it would be salutary to adopt this approach more widely in the moral responsibility debate, especially with regards to those arguments which (like the manipulation argument) appeal to intuitions. Many arguments in the literature, both skeptical and defensive of moral responsibility, depend on this strategy. As Sommers (2010: 200) notes, "the bottom line is that contemporary theories of free will and moral responsibility essentially involve appeals to intuitions about key principles and cases". This means that other arguments in the literature may fall prey to similar conceptual concerns as those which I claim afflict the manipulation argument. In fact, I think this

[221] Of course, I have myself at times in this chapter expressed things in terms of blame and praise, and blameworthiness and praiseworthiness. But this is only having *stipulated* in the first two chapters how I am interpreting these concepts. (To recap: this stipulation understands blameworthiness as a status which provides a normative reason to respond with blame, which is at minimum grounded in agential features deployed in subpar moral performance. It understands blame as any response had to an appraisal of, at minimum, agential features deployed in subpar moral performance). Moreover, the arguments of this chapter do not require that we accept this stipulation. In fact, they reveal how we need not invoke the concepts of blame and blameworthiness at all, neither my conception of them nor any other. Instead, we can more clearly articulate the skeptical considerations of the manipulation argument using the language of normative reasons for action.

is often the case. The approach I have taken with respect to the manipulation argument may therefore be instructive for how we might reframe other arguments too.

As a brief example, return to Doug Portmore's (2024) argument that the blameworthy deserve to suffer. In condensed form, this argument asks for our intuitions about a comparison between two possible worlds: one in which the blameworthy suffer, versus one where the non-blameworthy suffer. Portmore thinks the former world is clearly preferable to the latter. As he says, "[t]he idea, then, is merely that if someone has to suffer guilty feelings, it should be the blameworthy rather than the non-blameworthy" (2024: 811).

The problem here, I suggest, concerns how we are meant to understand *blameworthiness* for the purposes of this comparison. Here is the only specification which Portmore provides: the blameworthy are "those who have freely and knowingly done wrong" (2024: 795). So, the comparative question effectively becomes: do we share Portmore's intuition that if one is to suffer, it should be she who has freely and knowingly done wrong rather than she who has not?

But this begs another fundamental question. What does this agent's *freedom* with respect to her wrongdoing consist in? What kind of control does it endow her with? For example, is she free because she exercises reasons-responsiveness; or because she exhibits a poor quality of will; or because she possesses agent-causal libertarian freedom? Or something else?

How these questions are answered makes all the difference. The free will and moral responsibility debate has been occupied with addressing the existence, and normative significance, of *many* different varieties of freedom and control. If it remains ambiguous which of these varieties of freedom is possessed by Portmore's blameworthy agent, then it is not clear what his argument is able to establish. Even if one agrees that the blameworthy should suffer over the non-blameworthy, without reference to any specific agential features we make no progress towards establishing which agents this conclusion applies to, if it applies to any at all.[222]

[222] Indeed, I suggest that Portmore's description of the agent in question as blameworthy (and free), without any further specification of the agential features to which they owe this status, serves to *prejudice* our intuitions. It guarantees she is worthy of *some* blame response which the non-blameworthy is not. It is also natural to conceptually associate at least one conception of

These conceptual concerns are akin to those which I argue undermine Pereboom's manipulation argument. In my view we should therefore adopt the same methodology in reframing Portmore's argument. Its central question could instead be posed without invoking blameworthiness, or freedom, at all. We could ascribe *specific* agential features to the agent in question, providing us with a specific elaboration of what their freedom entails. For example, is it preferable that a fully reasons-responsive agent who knowingly does wrong suffers over one who does not? Reframing Portmore's argument in this way promises to deliver a less ambiguous conclusion: our intuitions tell us something concrete about the normative significance of the agential features in question. A similar approach might be ameliorative if applied to wider arguments in the moral responsibility literature that rely on concepts (such as blame, blameworthiness, and freedom) which threaten to be problematically ambiguous.

4.8. Concluding Remarks

After initially setting out the broader sweep of moral responsibility skepticism, the central focus of this chapter has been Pereboom's manipulation argument. This is the argument I have used to test for the kind of moral responsibility which we should understand 'basic desert' skepticism, as it has been presented, to in fact undermine.

I hope to have impressed how the machinery of Pereboom's argument – its premises, and the power of its four cases to elicit strong intuitions – is in very good working order. However, I raised concerns about the conceptual terms in which Pereboom frames the argument. He tests for intuitions about moral responsibility and basic desert. As I have established in earlier chapters, both are decidedly ambiguous and difficult concepts to pin down. It is difficult to know what conclusions we can confidently draw from testing intuitions about them, especially when it comes to folk intuitions.

I then argued that the manipulation argument's machinery can instead be more effectively trained on different and clearer targets, ones which it is naturally equipped to eliminate. It is not merely effective in undermining the worthiness – based on sheer

blameworthiness with the response of deserved guilt. But again, until we are more specific about the agential features which ground this blameworthiness, then I cannot see what the argument hopes to establish other than that we can conceive of *some* kind of blameworthiness, as yet unspecified, which would generate a particular reason to experience suffering.

compatibilist agential features – of those important kinds of harm-involving blame cited by the *Claim Forfeiture* and *Retributivism* accounts. The manipulation argument *also* impugns *Basic Reason* moral responsibility: it suggests that sheer (compatibilist) agential features cannot generate normative reasons in favour of actions of any kind. The veil of skepticism can therefore be thrown more widely.

I said in closing Chapter 1 that the idea that deployed agential features could not generate at least some reasons of blameworthiness might seem very unlikely. The findings of this chapter, however, deal a sizeable blow such a prediction. The conclusion I have arrived at is a strong one: the control imparted by our compatibilist agential features is real, but by itself – *sheerly* – it generates no reasons favouring actions of any kind. We should therefore cut compatibilist agential features down to size and accord them the same significance as we do our height – at least when it comes to their freestanding power to generate normative reasons for action.

These considerations realise the manipulation argument's untapped and undersold potential. They also allow us to swerve the unsettled conceptual difficulties which afflict the *Retributivism* and *Claim Forfeiture* conceptions of moral responsibility. We now derive from the manipulation argument a conceptually cleaner – and thus I think all the more powerful – brand of skepticism.

Even if one has misgivings about my interpretation of the manipulation argument, I also hope my arguments have modelled a salutary methodology which might be replicated elsewhere. My treatment of the manipulation argument – and the construction of my overarching conception of moral responsibility in Part One – reflects the value in recognising where important concepts are variably interpretable or ambiguous, and where we need to make precisely clear how we are wielding them if our arguments hope to be effective.

Conclusion

Here is a summary of what I take to be the central contributions of this thesis.

First, I diagnosed in moral responsibility scholarship the extreme, and at times confusing, variation in how certain very fundamental concepts are deployed and theorised about. Not least the concept of moral responsibility itself, but also blame (and praise) and blameworthiness (and praiseworthiness). I have sought to address this by developing a new overarching conception of moral responsibility. In the first place, this led me to articulate blameworthiness in terms of (at minimum) the deployment of agential features in moral underperformance. Then, I set out a complementary theory of blame (and by extension praise). This *Appraisal Theory* is strikingly thin, identifying blame simply in terms of what the blamer appraises. But I think this represents a novel way of thinking about blame, one which still allows the concept to demarcate a very significant category of responses (those which appraise the exercise of control in our moral underperformances).

I believe the resultant overarching conception of moral responsibility, which the *Appraisal Theory* completes, represents a useful contribution on its own terms. It makes good sense of the diverse range of moral responsibility theories in the literature. I hope it also reframes their familiar conceptual components in an illuminating way. Moreover, I think my conceptual framework offers an especially fruitful way of framing things for my subsequent analysis of basic desert moral responsibility in Part Two of the thesis. Here I opted to closely interrogate two of the most prominent accounts of basic desert: that offered by its skeptics (Pereboom and Caruso), and Menges' account. I ultimately argued that both accounts, whilst very valuable contributions, fail to reflect the true significance of the relevant skeptical considerations. These – at least those expressed in the manipulation argument – in fact recommend that we reconceive basic desert skepticism as basic *reason* skepticism. Our compatibilist agential features, when deployed in non-neutral moral performance, do not by themselves generate any reasons in favour of actions. They therefore lack the normative significance with which they are often credited (not least by the majority of moral responsibility theories). This finding is the second central contribution of the thesis.

Finally, as I have said, aside from the success of my arguments, I hope to have modelled a productive approach we can take to these matters. We do not necessarily need to go so far as constructing a full conception of moral responsibility. But the moral responsibility debate promises to benefit by theorists more patiently and explicitly set out what they understand its core concepts, such as blame and praise, to mean.

As likely is par for the course in any project such as this, I finish with a sense of some promising avenues left unexplored. Some of these result from recognised limitations in the reach of my arguments. Acknowledging these limitations simultaneously presents opportunities for further elaboration and refinement, so that we might build upon these initial findings. Here are a few possible directions for such future research:

- My arguments in support of my overarching conception of moral responsibility are predominantly *conceptual* in tenor. I have sought to establish credible conceptual parameters for blame or blameworthiness, rather than to defend any particular normative theory of these things (such as a reasons-responsiveness theory, say). But this conceptual groundwork is only really practically significant if it is a precursor to addressing these normative questions. I hope my conceptual clarifications, and framing in terms of deployed agential features, can help to inform further work on these normative questions (as they do, for instance, with regards to my discussion of basic desert).
- My arguments have centred upon blame and blameworthiness, and only derivatively on praise and praiseworthiness. I have indicated where I think my conception of blame and blameworthiness might apply, by extension and symmetrically, to the praising side of the coin. But I recognise this remains a mostly speculative claim, which further work would need to substantiate. (I have also focused mostly on other-directed blame, and generally assumed that my claims have extension to self-blame too. Subsequent work might pay closer attention to these distinctions, and examine whether my conceptual framework which treats them equivalently is sustainable.)
- In terms of the arguments of Part Two, I have only examined a single kind of skeptical strategy, one which weighs against compatibilist theories of moral responsibility. In order to build on this, we might take a similar approach to re-

examining skeptical arguments against *libertarian* accounts. Can these similarly be shown to discredit libertarian agential features as having *any* freestanding reason-giving significance, or do they reveal something else? Also, my refreshed version of the manipulation argument, like Pereboom's original version, is designed to test our intuitions. My arguments might be strengthened if we can show my intuitions are corroborated more widely, including in experimental testing with the folk.

- My discussion of *Basic Reason* moral responsibility (and skepticism thereof) strictly concerns implications for reasons for action. But I briefly gestured (in Footnote 209) at how I think the manipulation argument may yet weigh against a wider set of normative considerations. In particular, the fittingness of the reactive attitudes of anger and resentment. My sense is that the fittingness conditions of these emotions has commonly been misdescribed exclusively in terms of quality of will (an inheritance from P.F. Strawson). For now, these claims remain speculative, but I am keen to explore them further.

- The avenue I am perhaps most interested in exploring next concerns the *practical implications* of these ideas. When I first started thinking about these issues, I was especially interested in the practical implications of basic desert skepticism. Some initial work has already been done in this area (mostly on criminal reform, as discussed). However, it gradually became clear to me that before one could credibly address these practical questions, there was work to do upstream to clarify what basic desert even *is*. I therefore view my arguments in Part Two as contributing to this preliminary work. They leave us in a better place to investigate the practical implications. They also imply that these implications may be even more profoundly revisionary than the basic desert skeptics have envisaged. My arguments propose that our agential features are *entirely* devoid of freestanding reason-giving significance (at least when it comes to reasons for actions). If we take this seriously, it might substantially alter how we view and treat each other (and my sense is that this could be a very *positive* development). Also, whereas *basic desert* is never likely to ever catch on in common parlance, the relative conceptual straightforwardness of my arguments (as indicated in Chapter 4) perhaps means there may be more hope for these ideas to gain popular traction.

I will finish with a quote from Ursula Le Guin – partly just because I like it. But I think it also resonates with the thrust of at least some of this thesis' arguments. I firmly endorse the sentiment about freeing ourselves from the idea of deserving. But the latter arguments of this thesis suggest we might go one step further. We might also free our minds of the idea that our agency has the normative significance we so often accredit it with. The implications of us doing so remains an open question, but I am keen to continue exploring them.

> Speak not of what men deserve. For we each of us deserve everything, every luxury that was ever piled in the tombs of the dead kings, and we each of us deserve nothing, not a mouthful of bread in hunger. Have we not eaten while another starved? Will you punish us for that? Will you reward us for the virtue of starving while others ate? No man earns punishment, no man earns reward. Free your mind of the idea of deserving, the idea of earning, and you will begin to be able to think." Ursula Le Guin, *The Dispossessed*

Bibliography

Adams, Robert. 1985. “Involuntary Sins.” *The Philosophical Review*, 94(1): 3–31.

Arneson, Richard J. 2003. "The Smart Theory of Moral Responsibility and Desert," in S. Olsaretti (ed.), *Desert and Justice*: 233–258. Oxford University Press.

Arpaly, Nomy. 2006. *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton University Press.

Arpaly, Nomy and Schroeder, Timothy. 2014. *In Praise of Desire*. Oxford University Press.

Ayer, A.J. 1954. *Philosophical Essays*. Macmillan.

Baker, Lynne R. 2006. “Moral Responsibility without Libertarianism.” *Noûs*, 40(2): 307–330.

Balaguer, Mark. 2010. *Free Will as an Open Scientific Problem*. MIT Press, Bradford Books.

Basu, Rima. 2019. “What We Epistemically Owe to Each Other.” *Philosophical Studies,* 176(4): 915–31.

Bedau, Hugo Adam and Erin Kelly. 2015. “Punishment,” in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*. Available at: http://plato.stanford.edu/archives/fall2015/entries/punishment/ [Accessed 30th September 2024].

Bell, Macalester. 2013. “The Standing to Blame: A Critique,” in D. Justin Coates and Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*: 263–281. Oxford University Press.

Bennett, Christopher. 2002. “The Varieties of Retributive Experience.” *Philosophical Quarterly*, 52(207): 145–163.

Berman, Mitchell. 2008. “Punishment and Justification.” *Ethics*, 118(2): 258–290.

Berman, Mitchell. 2016. “Modest Retributivism,” in Kimberly Kessler Ferzan and Stephen Morse (eds.), *Legal, Moral, and Metaphysical Truths: The Philosophy of Michael S. Moore*: 35–48. Oxford University Press.

Bingeman, Emily. 2024. “The Risks (and Powers) of Praise,” in Santiago Amaya, David Shoemaker, and Manuel Vargas (eds.), *Oxford Studies in Agency and Responsibility Volume 8: Non-Ideal Agency and Responsibility*: 245–268. Oxford University Press.

Brink, David and Dana Nelkin. 2013. “Fairness and the Architecture of Responsibility,” in David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Vol. 1*: 284–313. Oxford University Press.

Calhoun, Cheshire. 1989. “Responsibility and Reproach.” *Ethics,* 99: 389–406.

Carlsmith, Kevin. M. 2006. “The Roles of Retribution and Utility in Determining Punishment.” *Journal of Experimental Social Psychology,* 42(4): 437–451.

Carlsmith, Kevin M. 2008. “On Justifying Punishment: The Discrepancy Between Words and Actions.” *Social Justice Research*, 21(2): 119–137.

Carlsmith, Kevin M. and John M. Darley. 2008. “Psychological Aspects of Retributive Justice.” *Advances in Experimental Social Psychology*, 40: 193–236.

Carlsmith, Kevin M., John M. Darley, and Paul H Robinson. 2002. “Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment.” *Journal of Personality and Social Psychology*, 83(2): 284–299.

Carlsson, Andreas Brekke. 2017. “Blameworthiness as Deserved Guilt.” *The Journal of Ethics*, 21: 89–115.

Caruso, Gregg D. 2012. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lexington Books.

Caruso, Gregg D. 2016. “Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model.” *Southwest Philosophy Review*, 32(1): 25–48.

Caruso, Gregg D. 2021. *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Cambridge University Press.

Caruso, Gregg D. and Stephen G. Morris. 2017. “Compatibilism and Retributivist Desert Moral Responsibility: On What is of Central Philosophical and Practical Importance.” *Erkenntnis,* 82(4): 837–855.

Caruso, Gregg D. and Derk Pereboom. 2022. *Moral Responsibility Reconsidered (Elements in Ethics)*. Cambridge University Press.

Clarke, Randolph. 1993. "Toward a Credible Agent-Casual Account of Free Will." *Noûs, 27*(2): 191–203.

Clarke, Randolph. 2003. *Libertarian Accounts of Free Will*. Oxford University Press.

Chisholm, Roderick M. 1976. *Person and Object: A Metaphysical Study*. Open Court.

Coates, D. Justin and Neal A. Tognazzini. 2012. "The Nature and Ethics of Blame." *Philosophy Compass,* 7(3): 197–207.

D'Arms, Justin, and Daniel Jacobson. 2000. "The Moralistic Fallacy: On the 'Appropriateness' of Emotions." *Philosophy and Phenomenological Research,* 61(1): 65–90.

D'Arms, Justin, and Daniel Jacobson. 2003. "The Significance of Recalcitrant Emotion (or, Anti-quasijudgmentalism)." *Royal Institute of Philosophy Supplements*, 52: 127–145.

Darley, John. M., Kevin M. Carlsmith, and Paul H. Robinson. 2000. "Incapacitation and Just Deserts as Motives for Punishment." *Law and Human Behavior*, 24(6), 659–683.

Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.

Davidson, Donald. 1980. *Essays on Actions and Events.* Clarendon Press.

Dennett, Daniel C. and Gregg D. Caruso. 2021. *Just Deserts*. Polity Press.

De Marco, Gabriel, and Taylor W. Cyr. 2024a. "Manipulation Cases in Free Will and Moral Responsibility, Part 1: Cases and Arguments." *Philosophy Compass*, 19(12): 1-11.

De Marco, Gabriel, and Taylor W. Cyr. 2024b. "Manipulation Cases in Free Will and Moral Responsibility, Part 2: Manipulator-focused Responses." *Philosophy Compass*, 19(12): 1-12.

De Marco, Gabriel, and Taylor W. Cyr. 2025. "Manipulation Cases in Free Will and Moral Responsibility, Part 3: Bypassing Responses." *Philosophy Compass*, 20(4): 1-12.

Doris, John. 2015. *Talking to Our Selves*. Oxford University Press.

Double, Richard. 1991. *The Non-Reality of Free Will*. Oxford University Press.

Dreyfus, Hubert L. 2002. "Intelligence Without Representation—Merleau-Ponty's Critique of Mental Representation," *Phenomenology and the Cognitive Sciences*, 1(4): 367–383.

Duff, R.A. 2010. "Blame, Moral Standing and the Legitimacy of the Criminal Trial." *Ratio*, 23(2): 123–140.

Duggan, A.P. 2018. "Moral Responsibility as Guiltworthiness." *Ethical Theory and Moral Practice*, 21: 291–309.

Ekstrom, Laura. 2000. *Free Will: A Philosophical Study*. Westview.

Elzein, Nadine. 2013. "Basic Desert, Conceptual Revision, and Moral Justification." *Philosophical Explorations*, 16(2): 212–225.

Feltz, Adam. 2013. "Pereboom and Premises: Asking the Right Questions in the Experimental Philosophy of Free Will." *Consciousness and Cognition,* 22(1): 53–63.

Feinberg, Joel. 1970. *Doing and Deserving*. Princeton University Press.

Fischer, John Martin. 2006. *My Way: Essays on Moral Responsibility*. Oxford University Press.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

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

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy*, 68(1): 5–20.

Frankfurt, Harry. 1988. "Identification and Wholeheartedness," in Harry Frankfurt (ed.), *The Importance of What We Care About: Philosophical Essays*: 159–176. Cambridge University Press.

Frankfurt, Harry. 1992. "The Faintest Passion." *Proceedings and Addresses of the APA,* 66(3): 5–16.

Franklin, Christopher. 2011. "Farewell to the Luck (and Mind) Argument." *Philosophical Studies,* 156(2): 199–230.

Fricker, Miranda. 2016. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs*, 50(1): 165–183.

Ginet, Carl. 1990. *On Action*. Cambridge University Press.

Griffith, Meghan. 2010. "Why Agent-caused Actions are not Lucky." *American Philosophical Quarterly*, 47(1): 43–56.

Haas, Daniel. 2013. "Merit, Fit, and Basic Desert." *Philosophical Explorations*, 16(2): 226–239.

Haji, Ishtiyaque. 1998. *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford University Press.

Haji, Ishtiyaque. 2009. *Incompatibilism's Allure: Principal Arguments for Incompatibilism*. Broadview Press.

Hart, H. L. A. 1968 [2008]. *Punishment and Responsibility: Essays in the Philosophy of Law (2nd Edition)*. Oxford University Press.

Hieronymi, Pamela. 2004. "The Force and Fairness of Blame." *Philosophical Perspectives*, 18(1): 115–148.

Hobbes, Thomas. 1651 [2021]. *Leviathan* (David Johnston, ed.). W.W. Norton & Company.

Holroyd, Jules. 2021. "Oppressive Praise." *Feminist Philosophy Quarterly,* 7(4): 1–26.

Holroyd, Jules. 2023. "Proleptic Praise: A Social Function Analysis." *Noûs*, 58(4): 905–926.

Hume, David. 1739 [1978]. *A Treatise of Human Nature*. Oxford University Press.

Hurley, Susan L. 2003. *Justice, Luck and Knowledge*. Harvard University Press.

Husak, Douglas. 2000. "Holistic Retributivism." *California Law Review*, 88: 991–1000.

Jefferson, Anneli. 2019. "Instrumentalism about Moral Responsibility Revisited." *The Philosophical Quarterly*, 69(276): 555–573.

Jefferson, Anneli. 2024. "Blaming the Dead." *European Journal of Philosophy*, 32(2): 548–559.

Jeppsson, Sofia. 2016. "Accountability, Answerability, and Freedom." *Social Theory and Practice*, 42(4): 681–705.

Jeppsson, Sofia. 2020a. "The Agential Perspective: A Hard-Line Reply to the Four-Case Manipulation Argument." *Philosophical Studies,* 177(7): 1935–1951.

Jeppsson, Sofia. 2020b. "Retributivism, Justification and Credence: The Epistemic Argument Revisited." *Neuroethics,* 14(2): 177–190.

Kahneman, Daniel, Schkade, David, and Sunstein, C. R. 1998. "Shared Outrage and Erratic Awards: The Psychology of Punitive Damages." *Journal of Risk and Uncertainty*, 16(1): 49–86.

Kane, Robert. 1985. *Free Will and Values*. State University of New York Press.

Kane, Robert. 1996. *The Significance of Free Will*. Oxford University Press.

Kane, Robert. 2007. "Libertarianism," in John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas (eds.), *Four Views on Free Will*: 5–43. Blackwell.

Kant, Immanuel. 1781 [1987]. *Critique of Pure Reason* (Paul Guyer and Allen Wood eds.). Cambridge University Press.

Kant, Immanuel. 1797 [1991]. *The Metaphysics of Morals* (Mary J. Gregor ed.). Cambridge University Press.

Kershnar, Stephen. 2000. "A Defense of Retributivism." *International Journal of Applied Philosophy*, 14(1): 97–117.

Kershnar, Stephen. 2001. *Desert, Retribution, and Torture*. University Press of America.

Kleinig, John. 1973. *Punishment and Desert*. Martinus Nijhoff Publishers.

Knobe, Joshua and Doris, John M. 2010. "Responsibilitys," in John Doris, *Moral Psychology Handbook*: 321–354. Oxford University Press.

Korsgaard, Christine. 1983. "Two Distinctions in Goodness." *Philosophical Review,* 92(2): 169–95.

Kubala, Robbie. 2024. "Aesthetic Blame." *Journal of the American Philosophical Association,* 10(4): 744–760.

Latham, Noa. 2004. "Determinism, Randomness, and Value." *Philosophical Topics,* 32(1-2): 153–167.

Lenman, James. 2006. "Compatibilism and Contractualism: The Possibility of Moral Responsibility." *Ethics*, 117(1): 7–31.

Levy, Neil. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.

Lippert-Rasmussen, Kasper. 2022. "Praising without Standing." *Journal of Ethics,* 26(2): 229–246.

List, Christian, Gregg Caruso and Cory Clark. "Free Will: Real or Illusion?" *The Philosopher*, 108(1): 61–80.

Lycan, William G. 1997. *Consciousness*. MIT Press.

Macnamara, Coleen. 2015. "Blame, Communication, and Morally Responsible Agency," in Randolph Clarke, Michael McKenna, and Angela M. Smith (eds.), *The Nature of Moral Responsibility: New Essays*: 211–236. Oxford University Press.

Matheson, Benjamin and Per-Erik Milam. 2021. "The Case against Non-Moral Blame," in Mark Timmons (ed.) *Oxford Studies in Normative Ethics, Volume 11*: 199–222. Oxford University Press.

McCormick, Kelly. 2022. *The Problem of Blame: Making Sense of Moral Anger*. Cambridge University Press.

McFatter, Robert M. 1978. Sentencing Strategies and Justice: Effects of Punishment Philosophy on Sentencing Decisions. *Journal of Personality and Social Psychology,* 36: 1490–1500.

McFatter, Robert M. 1982. “Purposes of Punishment: Effects of Utilities of Criminal Sanctions on Perceived Appropriateness.” *Journal of Applied Psychology,* 67: 255–267.

McGeer, Victoria. 2014. “PF Strawson’s Consequentialism,” in David Shoemaker and Neal Tognazzini (eds.), *Oxford Studies in Agency and Responsibility*: 64–92. Oxford University Press.

McGeer, Victoria. 2015. “Building a Better Theory of Responsibility.” *Philosophical Studies,* 172(10): 2635–2649.

McGeer, Victoria and Philip Pettit. 2015. “The Hard Problem of Responsibility,” in David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Vol 3*: 160–188. Oxford University Press.

McKenna, Michael. 2008. “A Hard-Line Reply to Pereboom’s Four-Case Manipulation Argument.” *Philosophy and Phenomenological Research,* 77(1): 142–159.

McKenna, Michael. 2012. *Conversation and Responsibility*. Oxford University Press.

McKenna, Michael. 2013. “Directed Blame and Conversation,” in D. Justin Coates and Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*: 119–143. Oxford University Press.

McKenna, Michael. 2014. “Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin.” *Philosophy and Phenomenological Research*, 89(2): 467–484.

McKenna, Michael. 2018. “Power, Social Inequities, and the Conversational Theory of Moral Responsibility,” in Katrina Hutchison, Catriona Mackenzie and Marina Oshana (eds.), *Social Dimensions of Moral Responsibility*: 28–58. Oxford University Press.

McKenna, Michael. 2020. “Punishment and the Value of Deserved Suffering.” *Public Affairs Quarterly,* 34(2): 97–123.

McKenna, Michael. 2024. *Responsibility and Desert*. Oxford University Press.

McKenna, Michael and Derk Pereboom. 2016. *Free Will: A Contemporary Induction*. Routledge.

Mele, Alfred R. 2005. "A Critique of Pereboom's 'Four-Case' Argument for Incompatibilism." *Analysis,* 65(1): 75–80.

Mele, Alfred. 2006. *Free Will and Luck*. Oxford University Press.

Mele, Alfred. 2008. "Manipulation, Compatibilism, and Moral Responsibility." *Journal of Ethics,* 12(304): 263–286.

Mele, Alfred. 2010. "Moral Responsibility for Actions: Epistemic and Freedom Conditions." *Philosophical Explorations,* (13)2: 101–111.

Menges, Leonhard. 2017. "The Emotion Account of Blame." *Philosophical Studies*, 174(1): 257–273.

Menges, Leonhard. 2022. "The Kind of Blame Skeptics Should Be Skeptical About." *Canadian Journal of Philosophy*, 51(6): 1–15.

Menges, Leonhard. 2023. "Responsibility, Free Will, and the Concept of Basic Desert." *Philosophical Studies*, 180: 615–636.

Milam, Per-Erik. 2021. "Get Smart: Outcomes, Influence, and Responsibility." *The Monist*, 104(4): 443–457.

Miller, David. 1976. *Social Justice*. Oxford University Press.

Montminy, Martin. 2021. "Defending the Epistemic Condition on Moral Responsibility." *Journal of Ethics and Social Philosophy,* 20(2): 168–187.

Moore, Michael. 2010. *Placing Blame: A Theory of the Criminal Law*. Oxford University Press.

Müller, Jean Moritz. 2021. "The Spontaneity of Emotion." *European Journal of Philosophy*, 29(4): 1060–1078.

Murray, Dylan, and Tania Lombrozo. 2017. "Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility." *Cognitive Science,* 41(2): 447–481.

Nadelhoffer, Thomas, Shahram Heshmati, David Kaplan and Shaun Nichols. 2013. “Folk Retributivism and the Communication Confound.” *Economics and Philosophy*, 29(2): 235–261.

Nahmias, Eddy, D. Justin Coates and Trevor Kvaran. 2007. “Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions.” *Midwest Studies in Philosophy,* 31(1): 214–242.

Nelkin, Dana. 2011. *Making Sense of Freedom and Responsibility*. Oxford University Press.

Nelkin, Dana. 2016. “Accountability and Desert.” *Journal of Ethics,* 20(1): 173–189.

Nichols, Shaun and Joshua Knobe. 2007. “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.” *Noûs*, 41(4), 663–685.

Nichols, Shaun, Mark Timmons, Theresa Lopez. 2014. “Using Experiments in Ethics – Ethical Conservatism and the Psychology of Moral Luck,” in Markus Christen, Carel van Schaik, Johannes Fischer, Markus Huppenbauer, Carmen Tanner (eds.), *Empirically Informed Ethics: Morality between Facts and Norms* (Library of Ethics and Applied Philosophy, vol 32). Springer, Cham.

Nozick, Robert. 1974. *Anarchy, State and Utopia*. Basic Books.

O’Connor, Timothy. 2009a. “Agent-Causal Power,” in Toby Handfield (ed.), *Dispositions and Causes*: 184–214. Oxford University Press.

O’Connor, Timothy. 2009b. “Degrees of Freedom.” *Philosophical Explorations,* 12(2): 119–125.

Parfit, Derek. 2011. *On What Matters, Vol. 1*. Oxford University Press.

Pereboom, Derk. 1995. “Determinism Al Dente.” *Noûs*, 29: 21–45.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge University Press.

Pereboom, Derk. 2008. “A Hard-line Reply to the Multiple-Case Manipulation Argument.” *Philosophy and Phenomenological Research,* 77(1): 160–170.

Pereboom, Derk. 2013. “Optimistic Skepticism about Free Will,” in Paul Russell and Oisin Deery (eds.), *The Philosophy of Free Will: Essential Readings from the Contemporary Debates*: 421. Oxford University Press.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pereboom, Derk. 2021a. “Undivided Forward-Looking Moral Responsibility,” *The Monist*, 104(4): 484–497.

Pereboom, Derk. 2021b. *Wrongdoing and the Moral Emotions*. Oxford University Press.

Pickard, Hanna. 2013. “Irrational Blame.” *Analysis*, 73(4): 613–626.

Portmore, Douglas. 2024. “Deserving to Suffer.” *The Journal of Ethics*, 28(4): 795–813.

Rachels, James (ed.). 1997. “What People Deserve.” *Can Ethics Provide Answers? And Other Essays in Moral Philosophy*: 175–197. Rowman and Littlefield Publishers.

Raz, Joseph. 1977. “Promises and Obligations,” in P.M.S. Hacker and J. Raz (eds.), *Law, Morality, and Society: Essays in Honor of H.L.A. Hart*: 210–228. Clarendon Press.

Repko Waller, Robyn. 2023. “Free Will and Self Expression: A Compatibilist Garden of Forking Paths.” *Philosophical Issues*, 33(1): 299–313.

Robinson, Paul H. 2008. *Distributive Principles of Criminal Law: Who Should Be Punished How Much*. Oxford University Press.

Roskies, Adina L. and Shaun Nichols. 2008. “Bringing Moral Responsibility Down to Earth.” *Journal of Philosophy,* 105(7): 371–388.

Rudy-Hiller, Fernando. 2022. “The Epistemic Condition for Moral Responsibility”, in Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy (Winter 2022 Edition).* Available at: https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/ [Accessed 14th March 2023].

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford University Press.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. Harvard University Press.

Scanlon, Thomas. 2008. *Moral Dimensions*. Harvard University Press.

Scanlon, Thomas. 2013. "Interpreting Blame," in D. Justin Coates & Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*: 84–99. Oxford University Press.

Schlick, Moritz. 1939. *Problems of Ethics.* Prentice-Hall.

Schlosser, Markus E. 2019a. "Agency," in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). Available at: https://plato.stanford.edu/archives/win2019/entries/agency [Accessed 18th December 2023].

Schlosser, Markus E. 2019b. "Dual-System Theory and the Role of Consciousness in Intentional Action," in Bernard Feltz, Marcus Missal and Andrew Sims (eds.), *Free Will, Causality and Neuroscience*: 35–56. Brill Editions.

Seim, Maria. 2019. "The Standing to Blame and Meddling." *Teorema*, 38: 7–26.

Shaw, Elizabeth. 2021. "The Epistemic Argument Against Retributivism." *Journal of Legal Philosophy,* 46(2): 155–160.

Shaw, Elizabeth, Derk Pereboom and Gregg D. Caruso (eds.). 2019. *Free Will Skepticism in Law and Society: Challenging Retributive Justice*. Cambridge University Press.

Sher, George. 2006. *In Praise of Blame*. Oxford University Press.

Shoemaker, David. 2007. "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community." *Ethics*, 118(1): 70–108.

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

Shoemaker, David. 2013. "Blame and Punishment," in D. Justin Coates and Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*: 100–118. Oxford University Press.

Shoemaker, David. 2015. *Responsibility from the Margins*. Oxford University Press.

Shoemaker, David. 2020. "Responsibility: the State of the Question: Fault Lines in the Foundations," *The Southern Journal of Philosophy*, 58(2): 205–237.

Shoemaker, David. 2022. "Empathic Control." *HUMANA.MENTE Journal of Philosophical Studies*, 15(42): 89–118.

Shoemaker, David, and Manuel Vargas. 2019. “Moral Torch Fishing: A Signaling Theory of Blame.” *Noûs*, 55(3): 581–602.

Sliwa, Paulina. 2019. “Reverse-Engineering Blame.” *Philosophical Perspectives*, 33(1): 200–219.

Smart, J.J.C. 1961. “Free-will, Praise and Blame.” *Mind,* 70(279): 291–306.

Smilansky, Saul. 2000. *Free Will and Illusion*. Oxford University Press.

Smith, Angela. 2005. “Responsibility for Attitudes: Activity and Passivity in Mental Life.” *Ethics,* 115(2): 236–271.

Smith, Angela M. 2007. “On Being Responsible and Holding Responsible.” *The Journal of Ethics*, 11(4): 465–484.

Smith, Angela. 2008. “Control, Responsibility and Moral Assessment.” *Philosophical Studies,* 138(3): 367–392.

Smith, Angela M. 2013. “Moral Blame and Moral Protest,” in D. Justin Coates & Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*: 27–48. Oxford University Press.

Smith, Michael. 2019. “Gary Watson: Strawsonian,” in Neal A. Tognazzini and D. Justin Coates (eds.), *Oxford Studies in Agency and Responsibility Volume 5: Themes from the Philosophy of Gary Watson*: 110–126. Oxford University Press.

Sommers, Tamler. 2007. “The Objective Attitude.” *Philosophical Quarterly*, 57(228): 321–341.

Sommers, Tamler. 2010. “Experimental Philosophy and Free Will.” *Philosophy Compass,* 5(2): 199–212.

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

Sripada, Chandra. 2016. “Self-Expression: A Deep Self Theory of Moral Responsibility.” *Philosophical Studies*, 173(5): 1203–1232.

Sripada, Chandra. 2017. “Frankfurt’s Unwilling and Willing Addicts.” *Mind*, 126(503): 781–815.

Stern, Lawrence. 1974. "Freedom, Blame, and Moral Community." *The Journal of Philosophy,* 71(3): 72–84.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford University Press.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies*, 75(1): 5–24.

Strawson, Peter. 1963. "Freedom and Resentment." *Proceedings of the British Academy,* 48: 187–211.

Tadros, Victor. 2009. "Poverty and Criminal Responsibility." *The Journal of Value 1120Inquiry,* 43(3): 391–413.

Talbert, Matthew. 2012. "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics*, 16(1): 89–109.

Talbert, Matthew. 2022. "Attributionist Theories of Moral Responsibility," in Dana Kay Nelkin and Derk Pereboom (eds.), *The Oxford Handbook of Moral Responsibility*: 50–70. Oxford University Press.

Talbert, Matthew. 2024. "Moral Responsibility," in Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/ [Accessed 23rd November 2024].

Taylor, Richard. 1974. *Metaphysics*. Prentice-Hall.

Telech, Daniel. 2021. "Praise as Moral Address," in David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Volume 7*: 154–181. Oxford University Press.

Telech, Daniel. 2022. "Praise." *Philosophy Compass,* 17(10): 1–19.

Telech, Daniel. 2024. "Standing to Praise." *European Journal of Philosophy,* 32(4): 1235–1254.

Tognazzini, Neal, and D. Justin Coates. 2021. "Blame," in Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*. Available at: https://plato.stanford.edu/archives/sum2021/entries/blame/ [Accessed 16th April 2024].

van Inwagen, Peter. 1975. "The Incompatibility of Free Will and Determinism." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 27(3): 185–199.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford University Press.

van Inwagen, Peter. 1989. "When Is the Will Free?" *Philosophical Perspectives*, 3: 399–422.

Vargas, Manuel. 2005. "The Trouble with Tracing." *Midwest Studies in Philosophy,* (29)1: 269–291.

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

Vicens, Leigh. 2022. "Free Will Skeptics Can Have Their Basic Desert and Eat It Too." *Journal of the American Philosophical Association,* 9(2): 355–366.

Vihvelin, Kadri. 2013. *Causes, Laws, & Free Will: Why Determinism Doesn't Matter*. Oxford University Press.

Vilhauer, Benjamin. 2009. "Free Will and Reasonable Doubt." *Philosophical Explorations*, 46(2), 131–140.

Vilhauer, Benjamin. 2012. "Taking Free Will Skepticism Seriously." *The Philosophical Quarterly*, 62(249): 833–852.

Walen, Alec. 2023. "Retributive Justice," in Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. Available at: https://plato.stanford.edu/archives/win2023/entries/justice-retributive/ [Accessed February 2024].

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

Wallace, R. Jay. 2010. "Hypocrisy, Moral Address, and the Equal Standing of Persons." *Philosophy and Public Affairs*, 38: 307–341.

Wallace, R. Jay. 2011. "Dispassionate Opprobrium: On Blame and the Reactive Sentiments," in R. Jay Wallace, Rahul Kumar and Samuel Freeman (eds.), *Reasons* *and Recognition: Essays on the Philosophy of T.M. Scanlon*: 348–372. Oxford University Press.

Wallace, R. Jay. 2019. *The Moral Nexus*. Princeton University Press.

Waller, Bruce N. 1990. *Freedom Without Responsibility*. Temple University Press.

Waller, Bruce N. 2011. *Against Moral Responsibility*. MIT Press.

Warmke, Brandon. 2011. "Moral Responsibility Invariantism." *Philosophia*, 39(1): 179–200.

Watson, Gary. 1975. "Free Agency." *The Journal of Philosophy*, 72(8): 205–220.

Watson, Gary. 1987 [2008]. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in Michael McKenna & Paul Russell (eds.), *Free Will and Reactive Attitudes*: 115–142. Ashgate.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics*, 24: 227–248.

Wolf, Susan. 1980. "Asymmetrical Freedom." *The Journal of Philosophy*, 77(3): 151–166.

Wolf, Susan. 1987. "Sanity and the Metaphysics of Responsibility," in Ferdinand Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*: 46–62. Cambridge University Press.

Wolf, Susan. 1990. *Freedom Within Reason*. Oxford University Press.

Wolf, Susan. 2011. "Blame, Italian Style," in R. Jay Wallace, Rahul Kumar and Samuel Freeman (eds.), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*: 332–347. Oxford University Press.

Zimmerman, Michael. 1988. *An Essay on Moral Responsibility*. Rowman and Littlefield.