Moral Agency

Timothy Nailer

B. Sci., B.A. Hons. (Philosophy), Grad. Dip. Ed.

Department of Philosophy

The University of Adelaide

Thesis submitted in fulfilment of the requirements for
the degree of Master of Philosophy

January 2022

ABSTRACT

While there is a vast philosophical literature exploring the conditions under which it is appropriate to hold individuals morally responsible for their actions, relatively little attention has been paid to the related question of which *kinds* of individuals merit these responsibility ascriptions. Under normal circumstances, typical adult human beings are held morally responsible for their behaviour but infants and nonhuman animals are not. In this thesis, I aim to account for this difference. That is, I aim to give an analysis of the concept of moral agency.

In Chapter One, I begin with a schema of moral agency, under which moral agents are characterised by the possession of certain abilities, enabling certain actions, for which certain responses are warranted. The literature on moral responsibility offers many ways of filling in these details, primarily by specifying the relevant agential abilities in terms of various responsibility conditions. My aim in this chapter is to offer a *basic* account, under which moral agents are characterised by the simplest possible abilities while still being appropriate targets of the relevant responses, such as praise and blame. I draw on the work of Nomy Arpaly, whose account of moral responsibility identifies the relevant ability as the ability to act out of good or ill will. I offer an analysis of this ability, under which basic moral agency is characterised by the ability to have desires about others' mental states.

In Chapter Two, I consider a range of alternative responsibility conditions offered by other philosophers, each of which appears to be necessary for moral agency. I argue that these conditions are not necessary for a basic account of moral agency.

In Chapter Three, I move beyond basic moral agency to offer a more restrictive account that aims to capture other important aspects of our moral lives: the practice of justification and our ability to improve our moral character. I claim these aspects are underpinned by the use of moral reasons to guide our behaviour, and in contrast to moral motivation, this guidance is characterised by the ability to have beliefs about the desirability of actions.

In Chapter Four, I aim to answer two questions: at what age do humans become moral agents, and are there any nonhuman animals who are moral agents. I draw on the work of Josef Perner, who offers a three-stage framework of the development of mental representation during childhood. I consider the conceptual coherence of this framework, its applicability to both accounts of moral agency developed in the thesis, and the empirical evidence that bears on the framework. As a result, I tentatively conclude that basic moral agency develops at around 18 months of age and may be present in a few species of nonhuman animal, whereas the more restrictive type of moral agency developed in Chapter Three develops no earlier than around 3.5 years of age and is restricted to human beings.

CONTENTS

THESIS DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

ACKNOWLEDGEMENTS

I am grateful to have had ample time and support to explore the ideas in this thesis. Especially as this is not my first attempt; I began a Ph.D. thesis about similar ideas in 2009. The decision to return to study was not an easy one, but thanks to the kind and persistent encouragement of Gareth Pritchard and Vesna Drapac, I once again saw the value in studying philosophy, and returned to study in 2017.

Having co-supervised my previous attempt, Garrett Cullity took a second chance on me and agreed to supervise the project. This thesis would not have been possible without his guidance. Throughout the project, his insight has made me a better thinker and a better writer. Since 2020, when Garrett left Adelaide for ANU, he has continued to show the same dedication, seeing my project through to the end. For all of this, I give my heartfelt thanks.

Upon Garrett's departure, Philip Gerrans took on the role of primary supervisor and had the unenviable task of handling the project's administrative paperwork during its last 18 months. He also came onboard as I was starting to write Chapter Four. This was a serendipitous occurrence, as his knowledge of the relevant empirical research greatly helped with the writing of this chapter.

My co-supervisor Gerard O'Brien also gave a helpful perspective, often as a counterpoint to the more analytical claims of the thesis. However, it is his guidance in dealing with the political challenges of the University that I am most grateful for.

Apart from my supervisory panel, I have benefitted from many discussions with staff and students in the Department of Philosophy. With apologies to those I have overlooked, thanks to Atheer Al-Khalfa, Kaz Bland, Brigitte Everett, Ben Fardell, Rob Farquharson, Jordi Fernandez, Alex Gabrielli, Ali Harwood, Claudia Ienco, Michael Lazarou, Chris Letheby, Michael Lopresto, Matthew Nestor, Greg O'Hair, Jon Opie, Paul Oppenheimer, Rónán Ruggles, Dook Shepherd, Di Stringer, Victoria Troitiño, James Vlachoulis, and Mathew Wenham.

My thesis owes a great deal to the work of Nomy Arpaly and Josef Perner. I was introduced to Perner's *Understanding the Representational Mind* by Jo Davis, while I was her research assistant at Adelaide Zoo. She gave me the opportunity to observe secondary representation in action in our interactions with Karta, a brilliant orangutan with a skill for deception. I later happened upon Arpaly's *Unprincipled Virtue* in the University library, no doubt because Garrett had requisitioned it, and took it with me to read on my honeymoon. These books were foundational in my understanding of moral agency, and this thesis is my synthesis of the ideas contained within them.

In June 2019, Nomy Arpaly visited Adelaide for our Departmental Seminar Series. I'd like to thank her for generously sharing her time and knowledge, and I'd like to thank those who made this possible, including Garrett Cullity for organising the series, as well as Tina Iankov and Lancy Xie for their administrative work in arranging the visit.

I received a great deal of help navigating the University's administrative systems from Kim Crawford, Tina Esca, Tina Iankov, Snezana Ilic, Ben McCann, Diane McInnes, Diana Reed, Courtney Sommer, and especially Dagmar Thiel and Tiziana Torresi. I offer kind thanks for all their help over the years.

I am also grateful to have received financial support from the School of Humanities, which contributed towards the purchase of an iPad, a laptop, and many books. I thank the School

for their contribution and encourage other research students within the School to avail themselves of these funds.

On the topic of funds, my candidature was punctuated by an unexpected career change and periods of financial instability. Kind thanks to Cheryl Baldwin, Demetri Bastiras, Cez Green, Helen Manning-Bennett, Tori Matthews, Sharon McAskill, Esther Speight, Megan Taylor, David Tyler, and Louise Woolford for offering advice, encouragement, and opportunities to help me change careers from administration back into teaching. Were it not for their help, I may have dropped out for a second time.

A suitable environment for writing is a necessity for any thesis project. I found mine at the Adelaide Central Market and the Coffee Barun Espresso Bar. Thanks to the staff at both locations for letting me work at their tables and for the much-appreciated food and coffee.

Balancing study with work and with parenting has not always been easy. I am fortunate, however, to have wonderful parents and in-laws, Sandy and Steve Nailer and Bert and Judi Rowe, who have supported my studies and who have never had a problem looking after my daughters when I've needed to write. I am forever grateful their support and for their presence in the girls' lives.

To my wife Michelle, you have been a source of strength from the very start. It's been incredibly rewarding to study philosophy while raising a family, which would have been impossible were it not for you. Before this, you gave me the confidence to pursue postgraduate study. Whenever it wavered, you were there to cheer me on and pick me up. I could not have done this without you.

Finally, to my daughters Emily and Madeleine, thank you for being you. I've sometimes joked that without you, this thesis would have been finished eight years earlier. That's not true. This is a thesis about moral agency, and you are moral agents. When I started for the first time, before you were born, I'd never seen a moral agent develop right before my eyes. But watching you two grow and learn and become better people, I've learned so much and I now know I couldn't have written this without you.

To my family – Shell, Emmy, and Maddie, – I dedicate this thesis to you.

INTRODUCTION

There is a sense in which we all know what moral agency is. When you or I do the wrong thing, it is often appropriate to hold us responsible for our wrongdoing. The same seems to be false of infants and animals. We do not hold them responsible when they act, at least not in the same way as when we act. This difference is the property of *moral agency*; we are moral agents and they are not.

This thesis sets out to explain this difference. Why is it appropriate to hold only some individuals responsible for their behaviour? What is the relevant difference that makes this appropriate?

The philosophical literature on these questions is vast but uneven. Much of this literature focuses on questions of *moral responsibility*: under which conditions is a moral agent responsible for their behaviour; which conditions justify or undermine our praise and blame of these agents? Less attention has been paid to the related question of which *kinds* of individuals can appropriately be held responsible, on which individuals are moral agents in the first place.

In this thesis, I draw on the extensive moral responsibility literature to answer a specific question about moral agency: which abilities do moral agents have? I do this by considering the responsibility conditions offered by theories of moral responsibility and observing that these are abilities of moral agents. If, for instance, a theory of moral responsibility claims that agents can only be held responsible for actions that are under their control, then this would imply that the ability to control one's actions is a necessary condition for moral agency.

In addition to responsibility conditions, accounts of moral responsibility differ in what kinds of responses they take to constitute the practice of holding agents responsible. These responsibility practices are varied, including attitudes (such as praise and blame), emotions (gratitude and resentment), and actions (reward and punishment). Different accounts of moral agency differ not only in which types of abilities are necessary for moral agency but in which types of responses are justified in virtue of how agents exercise these abilities.

This suggests that we can use accounts of moral responsibility to develop a general schema of a definition of moral agency, under which moral agents are characterised by the possession of certain *abilities*, which enable certain *actions*, the performance of which justifies certain *responses*. Depending on which abilities, actions, and responses one takes to be important, one can develop this schema into an account of moral agency in various ways.

The schema gives us a system for constructing an account of moral agency: identify a related set of abilities, actions, and responses. In particular, the relevant type of action should have *normative force*; all else being equal, moral agents ought to perform actions of this type. Once such actions are identified, we can then identify the abilities that enable these actions, and the appropriate responses to agents who either perform or fail to perform these actions.

In this thesis, I offer two complementary accounts of moral agency. The first aims to be maximally inclusive, demarcating agents who meet any plausible criterion for moral agency from those who are unambiguously not moral agents. The second account is more restrictive and aims to identify those moral agents who use moral reasons to guide their behaviour.

In Chapter One, I draw on the work of Nomy Arpaly to develop the first of these accounts. Among the more plausible accounts of moral responsibility, Arpaly's account is one of the most inclusive. It is so for three reasons.

First, it aims to justify responsibility for a simple kind of action: non-accidentally doing the right thing. As such, it does not require that moral agents be capable of more sophisticated behaviours, such as controlling their moral development or justifying their behaviour. Second, and because of this, it eschews many of the responsibility conditions characteristic of other accounts of moral responsibility in favour of the simpler ability to be motivated by moral reasons. And third, it seeks to justify a simpler response than many other accounts of moral responsibility. Rather than claiming that we ought to hold wrongdoers responsible by punishing them, it merely claims that we ought to adopt the negative attitude of blame toward such agents. Given this, Arpaly's account offers an excellent starting point for an inclusive account of moral agency.

I develop Arpaly's account by offering a detailed analysis of the ability to be motivated by moral reasons. I begin with commonsense psychology, which explains behaviour as arising from agents' beliefs and desires, and note that beliefs and desires are *content-bearing* mental states. From this, I infer that motivation by moral reasons involves desires with *moral content*.

This raises the issue of specifying this moral content. A particular outcome I wish to avoid is specifying moral content in such a way that I commit myself to a specific theory of normative ethics, such as claiming that a desire has moral content only if it is a desire about the maximisation of welfare. Given that no single theory of normative ethics is accepted by a majority of philosophers, I consider it a theoretical virtue for an inclusive account of moral agency to resist making such commitments. To this end, I offer an account of moral content that aims to be pluralist with respect to three central foundations of morality: *welfare*, *autonomy*, and *fairness*, such that a desire has moral content if it is about *either* others' welfare, others' autonomy, or fairness for others. The stipulation that the relevant desires must be about *others* aims to capture the intuition that morality is *other-regarding*.

I deepen this analysis by examining what distinguishes these types of moral content from non-moral content. In all three cases – of welfare, autonomy, and fairness – I argue that moral content is necessarily about *others' mental states*. Thus, the ability to have desires about others' mental states is a necessary condition for moral agency in its most inclusive sense.

In Chapter Two, I consider whether any other abilities are necessary for moral agency in this inclusive sense, and I argue that this is not the case. I do this by considering several accounts of moral responsibility that offer alternative responsibility conditions, and arguing that these responsibility conditions are not necessary for moral agency in the sense described above, although they may be necessary for other, less inclusive, accounts of moral agency. I consider the claims that responsibility depends on the following conditions: an agent's causal history, their knowledge of their actions, their endorsement of their actions, and their control over their actions.

I claim that these alternative responsibility conditions derive much of their intuitive force from the fact that they are taken to be necessary to enable other actions or to justify other responses that are taken to be relevant for moral responsibility. As such, my criticism of these alternative responsibility conditions is not aimed at showing these conditions to be unnecessary for these actions and responses, but that they are unnecessary for the simpler

actions and responses associated with the inclusive account of moral agency developed in the previous chapter.

However, the account of moral agency developed in the first two chapters, in virtue of its inclusivity, fails to capture important aspects of our moral lives, specifically, our practices of justification and improving our behaviour in light of moral reasons. I address this issue in Chapter Three by developing a complementary account of moral agency, more restrictive than the first, according to which moral agents are characterised by their ability to *guide* their behaviour by moral reasons. Because moral agency in the inclusive sense is a necessary precondition for the more flexible types of behaviour discussed in this chapter, I call these respective accounts *basic* and *flexible* moral agency.

Much of this account is focused on distinguishing this guidance from moral motivation, which characterises basic moral agency. The main contention is that guidance by moral reasons involves mental representation of those reasons, whereas mere moral motivation does not. Moreover, I argue that for moral reasons to guide behaviour, agents must recognise representational aspects of these reasons. Thus, guidance by moral reasons involves *metarepresentation*.

I argue that this ability to be guided by moral reasons enables agents to perform two actions that are of moral importance: justifying their actions and engaging in activities of moral improvement; that is, becoming a better person and helping others to do so too. In virtue of this, I claim that the appropriate response to the behaviour of flexible moral agents is to expect and to help them to engage in these activities.

Having developed two complementary accounts of moral agency in the first three chapters, I turn my attention in Chapter Four to the question of where moral agency is found in the world. Specifically, I am interested in two questions: at what age do humans become moral agents, and are there any nonhuman animals who are moral agents. Given the focus on applying accounts of moral agency to actual agents in this chapter, I draw on empirical literature in developmental and comparative psychology.

I begin with a distinction developed by the psychologist Josef Perner, between *metarepresentation* and the simpler representational abilities of *secondary representation* and *primary representation*. I note that while this distinction usefully demarcates distinct developmental stages in early childhood, it has trouble explaining desires as representational states. Given this, I modify the distinction to better accommodate desires while still maintaining its usefulness.

With a properly formed distinction between these three levels of representational ability, and having already established that flexible moral agency requires metarepresentation, I focus on basic moral agency and argue that this requires secondary representation but not metarepresentation. I do this by surveying the types of mental states that hold moral significance in considerations of welfare, autonomy, and fairness: *experiential states*, such as pleasure and pain; *intentional states*, such as beliefs and desires; and *emotions*, specifically reactive attitudes, such as gratitude and resentment, and I consider which level of representational ability is sufficient to have desires about these mental states. I argue that in none of these cases is primary representation sufficient but in at least some cases secondary representation is sufficient. Thus, I claim that secondary representation is necessary for basic moral agency.

Having argued that basic moral agency requires secondary representation, and that flexible moral agency requires metarepresentation, I then survey the psychological literature to determine the ages at which these abilities develop and whether they are present in nonhuman animals. I consider two abilities that are taken to be indicative of metarepresentation, *theory of mind* and *inhibitory executive function*, and two that are taken to be indicative of secondary representation, *pretend play* and *mirror self-recognition*. I concur with the claims that these abilities are indicative of the relevant level of representation and the (widespread but contested) view that secondary representation and metarepresentation develop in children at around 18 months and 3.5 years, respectively, with the caveat that there is individual variation in this timeline and that some circumstances may delay this development (such as autism or delayed language acquisition). I also concur with the view that secondary representation is limited to very few nonhuman species, and that metarepresentation is not present in nonhuman species. Thus, I claim that basic moral agency is limited to these nonhuman species and to humans over the age of 18 months, and that flexible moral agency is limited to humans over the age of 3.5 years.

CHAPTER ONE: BASIC MORAL AGENCY

What is Moral Agency?

In the months before I started writing this thesis, my eldest daughter would often ask me why she and her younger sister could do the same thing and only she would get in trouble. My answer was that her sister wouldn't get in trouble because she was just a baby. This wasn't very satisfying, to her or to me. What do babies lack that makes it inappropriate to hold them responsible for their actions? This thesis is my attempt to answer that question.

The simple answer is that babies lack *moral agency*. We are moral agents and they are not. But 'moral agency' is just a name, and without an analysis of the concept, this answer is no more satisfying than 'she's just a baby'. In this chapter, I offer an analysis of moral agency that explains why we only hold some people responsible for their actions.

But while much has been written about moral agency, most of this has been only indirectly about it, and surprisingly few philosophers have offered a definition. I like the structure of Tom Regan's definition:

"Moral agents are individuals who have a variety of sophisticated abilities, including in particular, the ability to bring impartial moral principles to bear on the determination of what, all considered, morally ought to be done and, having made this determination, to freely choose or fail to choose to act as morality, as they conceive it, requires. Because moral agents have these abilities, it is fair to hold them morally accountable for what they do, assuming that the circumstances of their acting as they do in a particular case do not dictate otherwise."[1]

That said, I think the definition is too narrow in several ways. Regan claims that moral agents are individuals. But it is at least conceivable that some moral agents may be groups instead of individuals.[2] It also strikes me that the use of impartial moral principles to determine an all-things-considered morally correct action does not capture our usual moral decision-making. This seems to be overthinking things. And nor do I think that agents are necessarily characterised by freely choosing to act in accordance with their conception of morality. Sceptics about freedom of the will often deny that agents have this ability but need not deny the existence of moral agents.[3] And finally, the definition emphasises the fairness of holding agents morally accountable, but it is an open question, firstly, whether the relevant relationship between agents' behaviour and the appropriate responses to their behaviour is one of fairness or of fittingness,[4] and secondly, whether the appropriate response is one of holding accountable or one of attributing praise or blame.[5]

That said, although I think Regan's definition is wrong with respect to these particulars, it nicely captures the general shape that such a definition ought to have. Moral agents are a type of *agent*, characterised by the possession of certain *abilities*, which enable certain *actions*, the performance of which justifies certain *responses* to them. Given this, I offer the following theory-neutral schema of a definition of moral agency:

---

[1] Regan (1983)

[2] See, for instance, List & Pettit (2011) and Isaacs (2011). Despite the plausibility of group agency, however, I will confine my focus to individuals as moral agents.

[3] See, for instance, Levy (2011) and Strawson (1994).

[4] Wallace (1994) and Arpaly (2002) take opposing views on this issue.

[5] In fairness, Regan's definition overlooks the accountability/attributability distinction because it predates the use of the distinction, which as far as I can tell, first appeared in Watson (1996).

"Moral agents are agents who have certain abilities, which enable the performance of certain actions. Because moral agents have these abilities, it is appropriate to respond to them in certain ways."

Of course, this schema is too broad to demarcate moral agents from other agents, but once we spell out the relevant abilities, actions, and responses, as Regan has done, we can use a definition of this form to make this demarcation. My task in this chapter is to spell out these abilities, actions, and responses.

Before I do, however, a few preliminary remarks are in order about agency in general. Part of my task in this chapter is to distinguish moral agents from other agents, which we might call 'non-moral agents', such as infants and chickens. But we may ask the further question about what distinguishes these agents from non-agents, such as rocks and molecules. Although the answer to this question is needed before we can properly analyse *moral* agency, a thorough examination of the question would take us far beyond the constraints of this thesis.

Thus, I offer here a reasonably uncontroversial account of agency. This account, known variously as *belief-desire psychology* or *common sense psychology*, like everything else in philosophy, is not universally accepted by philosophers but it does enjoy reasonably widespread acceptance.[6] On this account of agency, agents have two important types of mental states: *beliefs*, which aim to represent the word as it is, and *desires*, which represent goal states. Behaviour arises from the proper functioning of beliefs and desires. Thus, if I desire coffee and I believe that the café down the street serves coffee, then all else being equal, I will walk to the café. A key feature of both beliefs and desires is that they are *representations*,[7] which is to say that they are *about* something. In the example above, my desire is about coffee and my belief is about the café.

Thus, the account of moral agency I develop in this chapter is an account of agents with beliefs and desires and how they are different from other (non-moral) agents, also with beliefs and desires. I will argue that a key difference is that moral agents can have desires about some things that non-moral agents cannot have desires about.

The overall structure of this chapter is a division into five sections. In this, the first section, I have given a broad schema of a definition of moral agency and made some preliminary remarks below about agency in general. In the second and third sections, I discuss, respectively, the appropriate *responses* one may have toward moral agents, and the *actions* characteristic of moral agency. I argue that moral agents in the most inclusive sense warrant *praise* for acting out of *good will*, and they warrant *blame* for acting out of *ill will* or out of *insufficient good will*. My analysis of the *abilities* necessary for moral agency is discussed in the fourth section. In it, I argue that the key ability for moral agency is to have *desires with the relevant content*, and that this content is *others' mental states*. In the final section, I observe that this account is so inclusive as to potentially count toddlers and some nonhuman animals as moral agents. This is discussed in greater detail in Chapter Four.

---

[6] Some philosophers are sceptical about belief-desire psychology: see, for instance, Churchland (1988). However, most philosophers working on agency from David Hume (1738/1985) onwards seem to accept some version of belief-desire psychology. Michael Smith, for instance, states that this is the "*standard picture of human psychology*" (Smith 1994, p.7).

[7] Crane (2003)

A key feature of my approach is that I develop my account of moral agency from Nomy Arpaly's account of *moral responsibility*. There are several reasons for this. The first is that while there are relatively few accounts of moral agency specifically, there are very many accounts of moral responsibility. This is fortunate, as there is a straightforward relationship between moral agency and moral responsibility: *only* moral agents can be held morally responsible for their actions, and *all* moral agents can potentially be held morally responsible, assuming they act in ways that warrant this.[8]

This characterisation is meant to include normal adult human beings (the paradigmatic moral agents) even if they avoid altogether acting in ways that justify their being held morally responsible, but the characterisation means to exclude individuals, such as infants, who are simply incapable of acting in such ways.

Importantly, given that an account of moral responsibility specifies the conditions under which it is appropriate to hold agents morally responsible, these *responsibility conditions* can be recast as *agential abilities*. For instance, consider an account of moral responsibility according to which it is appropriate to blame a moral agent for an act of wrongdoing only if that agent could have done otherwise.[9] On such an account, one could only qualify as a moral agent if one were able to act (in at least some cases) in ways other than the way one did. Many responsibility conditions can be similarly recast in this way. Thus, it is possible to derive accounts of moral agency from accounts of moral responsibility by recasting these responsibility conditions as agential abilities. We can summarise this relationship between moral agency and moral responsibility as follows: *moral agents are those individuals with the ability to meet the relevant responsibility conditions*.

Of course, different accounts of moral responsibility offer different responsibility conditions, and thereby lead to different accounts of moral agency, some of which I will discuss in more detail in Chapter Two. The reason I have chosen Arpaly's account, though, is because it is inclusive. For instance, it is a widespread, but not universal view among philosophers working on moral psychology that psychopaths are not morally responsible, or at least not as responsible as non-psychopaths.[10] I must admit that my intuitions pull me in the opposite direction here. Psychopaths, of the type that appear in the moral psychology literature, strike me as bad people, who unlike babies, should be held morally responsible.

One could, of course, develop an account of moral agency in which psychopaths are exempt from responsibility in virtue of their specific psychological shortcomings.[11] Such

[8] Fischer (1987) and Eshleman (2019) have also noted this relationship between moral agency and moral responsibility.

[9] This was the dominant account of moral responsibility throughout much of the mid-twentieth century, during which a main point of contention was how to understand the phrase "could have done otherwise", with compatibilists preferring an interpretation that was compatible with determinism and incompatibilists preferring an interpretation that was not. See, for instance, Hobart (1934), Austin (1956), Schlick (1963), and Smart (in Smart & Williams (1973)). Work in the 1960s and 70s did much to change this view. Landmark papers include Frankfurt (1969), which argued against this as a responsibility condition; Strawson (1962), Frankfurt (1971), and Watson (1975), which offered accounts with other plausible responsibility conditions; and van Inwagen (1975), which argued that the compatibilist interpretation of "could have done otherwise" is not compatible with determinism.

[10] See, for instance Kennett & Fine (2008) and McGeer (2008), who develop conflicting accounts of moral responsibility, based in part on the fact that while they both accept that psychopaths have mitigated responsibility, they disagree over why this is the case.

[11] In addition to Kennett & Fine (2008) and McGeer (2008), I take Wolf (1990) to offer one such account.

an account may, for instance, hold that retributive punishment is appropriate only for non-psychopaths. However, I do not aim to give such an account. In offering my account of moral agency, my aim is to identify the most basic criteria for moral agency, such that any individual who fails to satisfy these criteria is not a candidate for moral agency under any plausible description. It is relatively uncontroversial that infants are not moral agents, whereas the moral agency of psychopaths is questionable. Thus, on a sufficiently inclusive account of moral agency, psychopaths ought to count as moral agents.

As we shall see, Arpaly offers an account of moral responsibility that is characterised by a very basic responsibility condition, and which thus can be developed into a very inclusive account of moral agency. Of course, if one thought that this account were too inclusive, one could develop some other account of moral responsibility into a more restrictive account of moral agency. I shall offer one such restrictive account in Chapter Three. For now, though, we shall consider appropriate responses to the behaviour of moral agents.

Praise and Blame as Evaluation

Theories of moral responsibility aim to determine the conditions under which it is appropriate to hold agents responsible for their actions. But different accounts of moral responsibility differ in what they take 'holding responsible' to mean. Consider *blaming*, a paradigm case of holding an agent responsible. Depending on who one is talking to, blame can refer to any of the following responses:

1. The mere evaluation of an agent's badness
2. An emotional response, such as indignation
3. The expression of an emotional response, such as a look of disdain
4. An utterance, such as "the broken window is your fault"
5. Social censure arising as the result of such expressions and utterances, either intentionally or unintentionally
6. Punishment proper, wherein the blameworthy individual is seen as deserving rebuke, and punished accordingly, on the basis of his blameworthiness

I've listed these roughly in order of the extent of their effect on the blamed individual but it is clear that they are each distinct phenomena and it is reasonable to believe that justifications for blame in some of the senses above do not necessarily justify blame in some of the other senses. For instance, one might justifiably feel indignant upon being wronged but conclude that the bar for punishment is not met.

While blame can refer to any of these responses, I will speak of blame as the evaluation of an agent as bad, and I will speak of blameworthiness as the agent's badness itself, such that it warrants this evaluation. In so doing, I recognise that the other, more externally-focused, senses of 'blame' are possible contenders for the 'correct' sense, but I do so for two reasons.

First, it seems to be true that the justification for blame-as-evaluation is either unrelated to or prior to the justification of blame-as-action. Consider the standard justifications for punishment: deterrence, rehabilitation, protection, and retribution.[12]

The first three are consequentialist justifications; punishment for deterrence, rehabilitation, and protection is justified because this produces good consequences. A traditional criticism of consequentialist justifications of punishment is that they are

[12] Tognazzini & Coates (2021)

unrelated to the target's *actual* blameworthiness. One could justifiably punish (or blame-as-action) innocent parties on such grounds.[13]

Retribution, by contrast, is justified on the grounds of the target's actual blameworthiness.[14] The 'blameworthiness' in this sense isn't merely the brute fact that the target deserves punishment (this would be tautologous and thereby offer no justification for punishment) but rather an evaluation of the agent (blame-as-evaluation), such that they meet certain criteria justifying their punishment. Blame-as-evaluation is thereby prior to retributive justifications for punishment (and blame-as-action in general).

The second reason I focus on blame-as-evaluation is that I am not primarily concerned with analysis of the concepts of blame or blameworthiness, but with the boundary conditions for moral agency. As the most basic sense of 'blame', blame-as-evaluation offers a demarcation criterion between non-moral agents, such as infants, and potential moral agents in the most minimal sense, such as children at the earliest appropriate point in their psychological development.

Blame in this sense is a negative evaluation of an agent. One may also make a positive evaluation. Such positive evaluations are typically referred to as *praise*, although this terminology connotes some asymmetries between praise and blame, such as the fact that it seems strange to praise someone silently, whereas blame is often left unspoken. Nonetheless, I will use the terms 'praise' and 'blame' to refer to evaluations of agents as, respectively, morally good and bad.

While I am concerned with evaluations of praiseworthiness and blameworthiness, it is also instructive to discuss how these evaluations relate to other types of moral evaluation. In particular, we tend to make moral evaluations of three types of things: states of affairs, actions, and agents.

A description of a state of affairs is a description of the way things are at a certain place and time. The state of affairs in which a child is happy, well fed, and well looked after is a good one. The state of affairs in which a child is unhappy, malnourished, and neglected is a bad one.

Actions refer to things that agents *do*, things that affect states of affairs. Murder is an action. In the context of normative ethics, actions also refer to omissions, things agents *fail to do*.[15] Refraining from murder is an action. All else being equal, murder is a bad action and refraining from murder is a good action. In addition to evaluating actions as good or bad, we also *prescribe* actions as right or wrong, as permissible or forbidden, as obligatory or optional, and very occasionally, as supererogatory or suberogatory. These prescriptions combine an evaluation of an action (as good or bad) with guidance on whether one ought to perform the action.

When evaluating agents, it is typically the case that a good agent is one who, all else being equal, performs good actions, whereas a bad agent is one who performs bad actions. Now this characterisation is, admittedly, incomplete, and I will fill in some of these details in the

[13] See, for instance McCloskey (1957).
[14] Walen (2021)
[15] Although there is dispute over the relative importance of actions and omissions (see, for instance, Foot (1967) and Tooley (1972)), it is uncontroversial that at least some omissions are considered wrong. Child neglect is a paradigm example.

following section. But at its most basic, the goodness of an agent is directly related to the goodness of their actions.

Quality of Will

At a first glance, the fundamental action characteristic of moral agency seems to be acting morally. By this, I mean acting in ways that are morally good, but I also mean acting in ways that are morally bad. That is, when I use the phrase 'acting morally', I am referring to actions that are appropriate targets of moral evaluation.

This section aims to give an account of acting morally such that doing so makes one praiseworthy or blameworthy in the sense described in the previous section. In it, I give a summary of Arpaly's views, which I take to be largely correct, on how to understand 'acting morally' such that it justifies praise and blame.

It is important to note that while I take Arpaly's account to be generally correct with respect to the general conditions for praiseworthiness and blameworthiness, she and I come to quite different conclusions regarding the presence of moral agency in children and animals. In short, she has claimed that children become moral agents around the time they correctly use the phrase "it's not fair"[16] and that animals are not moral agents.[17] I claim that moral agency emerges much earlier in childhood, at around 18 months of age, and that a few species of nonhuman animal may also be moral agents. These are controversial claims, which I discuss at length in Chapter Four, and are a result of our different views about which background conditions are necessary for moral motivation. I briefly discuss these differences toward the end of this chapter.

As we shall see, Arpaly treats praise and blame differently in virtue of the observation that it often seems appropriate to blame individuals for unintentional wrongdoing, such as forgetting to honour a promise, but it does not seem appropriate to praise agents for unintentional rightdoing, such as donating to charity for the sole purpose of claiming the tax deduction. Given this, I will consider praiseworthy action first, and then blameworthy action.

On Arpaly's account, an agent's praiseworthiness is constituted by their having done the right thing for the right reasons.[18] Three points are worth noting here:

First, praiseworthiness isn't constituted by merely doing the right thing, because one can do the right thing for the wrong reasons, such as donating in order to claim a tax deduction, or for no reason at all.[19]

Second, one's praiseworthiness is not merely due to a causal relationship between one's intentions and the outcomes of one's actions. For instance, Arpaly asks us to imagine a world in which the profit motive reliably produced desirable outcomes, such as the world imagined by advocates of market solutions for societal problems.[20] Even in this world, it seems that the agent who donates to Oxfam in order to minimise her tax obligations does not merit the same praise as the agent who does so in order to improve the world.

Rather, it seems that the agent who does the right thing for the right reasons merits praise because her reasons for action are the same as the reasons that make her act a good one

[16] Arpaly (personal communication – 2019)
[17] Arpaly (2002), p. 125; pp. 144-148
[18] *Ibid.*, p. 70
[19] *Ibid.*, p. 69
[20] *Ibid.*

to perform. Donating to charity is morally desirable because it improves the world. The agent who is motivated by this fact thereby merits praise for her action.

Third, agents who merely do what *they believe is right* do not merit praise if their beliefs are at odds with what is actually right. This is the case even if they perform morally desirable actions on the basis of their belief. If one donates to charity because one believes that minimising one's tax obligations is the right thing to do, one simply is mistaken, not praiseworthy. The fact that one improves the world by doing so is merely a happy accident, not something that warrants praise.

In short, for an agent to be morally praiseworthy is for her to have performed a right act for the reason that makes it right. The reason for which she acts is identical to the reason for which it is the right act to perform.[21]

Arpaly often refers to agents acting in these ways – ways that merit praise – as acting from *good will*. She contrasts this with *ill will*, which is the inverse of good will. Agents acting out of ill will are blameworthy in the same way that praiseworthy agents are praiseworthy.

Thus, for an agent to be morally blameworthy is for her to have performed a wrong act for the reason that makes it wrong. The reason for which she acts is identical to the reason for which it is the wrong act to perform. [22]

This analysis of blameworthiness applies well enough to actions performed out of spite or cruelty – that is, for actions performed out of ill will. But blameworthiness is more complex that praiseworthiness on Arpaly's analysis because there is an additional way of being blameworthy: acting out of *insufficient good will*.

Agents who act out of insufficient good will aren't cruel or spiteful, but their actions demonstrate an insensitivity to relevant moral considerations. A husband who upsets his wife because he fails to consider her feelings is blameworthy but not in the same way as a husband who does so because he enjoys seeing his wife upset. The first husband is acting from insufficient good will while the second is acting from ill will.

Agents who act out of insufficient good will, like those who act out of ill will, are blameworthy for doing the wrong thing. However, instead of acting for the reasons that make the action wrong, agents who act out of a lack of good will are blameworthy because they are *capable* of acting for the right reasons but *fail to do so*.

Thus, the ability necessary for moral agency on this account is the ability to act out of good or ill will. In the following section I will offer an analysis of this ability, such that we can (a) distinguish the morally salient motivations described by *good will* and *ill will*, such as compassion and spite, from morally neutral motivations, such as the profit motive, and (b) distinguish agents who are blameworthy for acting out of insufficient good will from agents who are simply *incapable* of acting out of good will, such as very young infants, and thereby do not merit blame for their actions.

Desires with Moral Content

It is important to get clear on what is meant by the ability to act out of good or ill will, since it is this ability that distinguishes moral agents from non-moral agents. As suggested

[21] *Ibid.*
[22] *Ibid.*

above, the exercise of this ability – acting out of good or ill will – involves an identity relation between one's reason for action and the evaluative properties of this action: acting out of good will requires that one's reason for action is identical to the good-making features of this action, and acting out of ill will requires that one's reason for action is identical to the bad-making features of this action. But how do we cash out this identity?

Arpaly comes closest to giving a complete description of this in the following passage:

"I take good will to be the same as […] responsiveness to moral reasons. I take a person to be responsive to moral reasons to the extent that she wants noninstrumentally to take courses of action that have those features that are (whether or not she describes them this way) [good]-making and not to take courses of action that have those features that are (whether or not she describes them this way) [bad]-making features. If good will – the motive(s) from which praiseworthy actions stem – is responsiveness to moral reasons, deficiency in good will is insufficient responsiveness to moral reasons, obliviousness or indifference to morally relevant factors, and ill will is responsiveness to sinister reasons – reasons for which it is never moral to act, reasons that, in their essence, conflict with morality."[23]

In this passage, responsiveness to moral reasons – acting for good reasons – is identified as motives from which praiseworthy actions stem. I think this is correct, but it raises the question of how to explain the identity relation between a motive – a desire – and the right-making features of an act.

The answer, I think, lies in the *representational* nature of desires. As described above, desires aim to represent the world not as it is but as it could be. An agent's desire to, say, help those in need will motivate one to help those in need. This desire represents the world as it could be – a world in which the needy are helped – and motivates one to change the actual world so as to bring it into line with this representation.

The identity between the good making features of an act and the reasons for which one acts are cashed out in the representational content of the relevant desires, such that the right making features are represented by the desire, which motivates the agent to act according to the content of the representation. Thus, one's desire to help those in need, in virtue of representing the right making features of donating to charity and by motivating one to do so, provide the connection between one's reason and one's action.

It should be noted that desires are not the only representational mental states involved in the production of actions. Belief-desire psychology aims to explain action in terms of a belief-desire pair.[24] For instance, one may desire to help those in need and one may believe that by donating to charity, one will help those in need. This belief-desire pair produces the action of donating to charity.

By contrast with desires, beliefs aim to represent the world as it is. If I believe that the sky is blue, then I have a mental representation of the sky as being blue. The agent's belief that donating to charity will help those in need is a representation of the causal relationship between donating to charity and helping the needy.

But it is one's desires, rather than beliefs, that determine one's quality of will. Some reflection will show that this is the case. Consider the representational contents of our agent's belief – a causal relationship between donating to charity and helping the needy – and their desire - a world in which the needy are helped. Suppose now that our agent

[23] *Ibid.*
[24] Davidson (1963)

instead believed that *refraining* from donating to charity would help the needy. In this case, she would refrain from donating to charity and thereby fail to be praiseworthy, not due to a lack of good will but because her false belief caused her good will to be misdirected.

Compare this with the inverse situation in which our agent's belief accurately represents a causal relationship between donating to charity and helping the needy, and desire that represents a world in which the needy are *not* helped. Our agent's behaviour in this situation would be the same as in the previous case: she would refrain from donating to charity. But her reason would be to *prevent* the poor from being helped. She also fails to be praiseworthy in this situation, but not merely because she fails to do the wrong thing, but also because her desire is not constitutive of a good will.

Another point to note is that Arpaly speaks specifically of *noninstrumentally* desiring to do the right thing. This refers to a distinction between *instrumental* and noninstrumental, or *intrinsic*, desires. The difference is that intrinsic desires are foundational, whereas instrumental desires are derived from belief-desire pairs. For instance, a desire to help the needy, when combined with a belief that donating to charity would achieve this, could generate an instrumental desire to donate to charity. However, a desire to minimise one's tax obligations, when combined with a belief that donating to charity would be an effective means of doing so, would also create an instrumental desire to donate to charity.

For this reason, it is the content of intrinsic desires that determines the quality of one's will and thereby whether an agent will be praiseworthy, blameworthy, or neither in acting on that desire. As mentioned above, the content of our desires provides the crucial link between the right making features of an act and the reasons for which one acts, as this content both *represents* the right making features and *motivates* agents to bring about these features.

Given that the reasons for which we act are specified by the representational content of our intrinsic desires, we can use this content to distinguish between *moral* reasons and other reasons. My desire for a coffee represents my having a coffee but there's nothing moral about this desire. If I were to act upon it, I wouldn't be acting for a moral reason. If one acted on one's desire to help the needy, by contrast, then one would be acting for a moral reason. The difference between the two lies in the content of the desires.

What kind of content, then, distinguishes moral reasons from non-moral reasons? We must be careful in answering this question because, on the face of it, this task seems to be the same as identifying the correct normative theory. For instance, if the correct normative theory is utilitarianism, then this would imply that the right kind of content is about maximising welfare. Or if the correct theory is Kantian deontology, then this would imply that the right kind of content is about treating persons as ends in themselves or acting on universalisible maxims.

But a good account of moral agency, and by extension the kind of content that constitutes moral reasons, should be neutral with respect to normative theories. If my account defines the right kind of content in a strongly Kantian way, then it is unlikely to convince those who have independent reasons to prefer utilitarianism.

For this reason, I think it's better to think of the task of identifying the right kind of content as answering a different, metaethical, question: what does a theory have to be

talking about to be a normative theory of morality?[25] While utilitarian and Kantian theories have their differences, there is a sense in which they are both talking about the same thing.

At a first glance, this 'same thing' seems to be how we ought to act. That said, there are theories of how we ought to act that are distinctly non-moral.[26] A choreographer's instructions to her dancers tell them how they ought to act while performing on stage, but these instructions don't constitute a normative theory of morality.

A better place to begin is with other-regardingness. Normative moral theories are theories that tell us how we should act toward others.[27] Therefore, the right kind of content is content that is other-regarding. My desire for a coffee has content that is about either a coffee or about me having a coffee. My desire to help those in need has content that is about those in need. My second desire is other-regarding, whereas my first is not.

An obvious objection to this line of thinking is that both Kantian and utilitarian theories endorse some actions that are not other-regarding. One can make Kantian arguments against drunkenness on the grounds that being drunk involves treating oneself as a mere means to gain pleasure at the expense of one's autonomy, which is impaired by alcohol. And one could make a utilitarian argument in favour of euthanasia in cases where the grief suffered by one's loved ones is outweighed by the relief to one's own suffering.

I think this line of objection is mistaken. Both utilitarianism and Kantian ethics *universalise* certain kinds of self-interested considerations, and it strikes me that it is this universalisation that makes them theories of ethics, rather than mere strategy or prudence.

Peter Singer claims that the basis for his utilitarianism is the principle of the equal consideration of interests.[28] That is, we can't treat our own interests as more important than those of others' merely because they are our own. We are to treat similar interests similarly, regardless of whose interests they are.

Likewise, the Kantian categorical imperative directs us to "act only in accordance with that maxim through which [one] can at the same time will that it become a universal law".[29] To do otherwise would be to make an exception of ourselves, an irrational act given that we are no more or less important than other members of the moral community. As with utilitarianism, we are to treat similar cases similarly.

With this understanding, the cases in which utilitarians and Kantians advocate acting in one's own best interests are best seen as specific applications of general principles, no different than other cases in which we act in others' best interests. The important point – the thing that distinguishes these theories from prudential norms – is that they are necessarily concerned with other people.

One might object at this point that the kind of enlightened self-interest promoted by egoists, as well as by related theories, such as Hobbesian contractarianism, does not disqualify these theories as moral theories.[30] I think this is mistaken but this does not matter for my purposes here. Whether moral theories are intrinsically or instrumentally other-regarding, the kinds of desires that constitute moral reasons will still contain content about others; the desires will still be other-regarding. Whether I keep my promise for

[25] Thanks to my supervisor Garrett Cullity for this suggestion.
[26] Gert & Gert (2020)
[27] *Ibid.*
[28] Singer (1979)
[29] Kant (1785/2002)
[30] Hobbes (1651/2017)

prudential or for intrinsically other-regarding reasons, my desire to do so will refer to the person to whom I am keeping the promise, and in this sense, be other-regarding.

Of course, if one were to claim that moral theories must be *noninstrumentally other-regarding*, then this would disqualify egoism and contractarianism as genuine moral theories. This is, in fact, the position I hold. But my point is that one *need not* discount these theories as genuine moral theories in order to endorse the claim that moral theories are necessarily other-regarding.

My contention, then, is that if someone cannot have other-regarding desires – if their psychology is limited in such a way that they cannot represent others in the right way – then they are not moral agents in even the most minimal sense. They are not good or bad, not in the sense that these evaluations are applied to moral agents. They are not the kind of things to which these evaluations even apply. Inanimate objects – rocks, for instance – are not moral agents. They don't have a psychology at all, much less the kind of psychology that can represent others in the right way.

With this in mind, let's consider the simplest possible desires to determine whether the agents who have them could be considered moral agents.

The thermostat is a simple device that can be described as having belief-like and desire-like states.[31] It has a sensor that represents the current temperature of the room. It has a mechanism that represents the 'desired' temperature of the room. And it has a mechanism that changes the temperature of the room if the difference between the current and desired temperature reaches a certain threshold.

At this point, the thermostat couldn't be described as having other-regarding desires except if we consider the room in which it operates as an 'other'. Of course, we can tweak the example so that this is the case. Imagine the thermostat is placed in an incubator, directly connected to an egg that requires a specific temperature in order to hatch. In this case, the thermostat 'desires' a temperature that is conducive to the hatchling's wellbeing, although it does not represent the hatchling directly.

I am not inclined to think of this thermostat as a moral agent but not because it doesn't *directly* represent the hatchling. We can further imagine that the thermostat's sensor *is* directly attached to the animal inside the egg and directly represents its body temperature. Even in this case, it seems to me that the thermostat isn't a moral agent.

My intuition here is that body temperature is the wrong target. Certainly, I can care about others' body temperature and this can drive me to act in what I consider a moral way, such as when my daughter has a fever. But this seems different to the non-moral concern I have when I adjust the oven temperature so as not to burn the roast pork. The 'concern' of the incubator thermostat seems more like the latter than the former.

The difference, I think, is that when I'm concerned about my daughter's temperature, I'm ultimately concerned about her welfare. I'm not concerned about my roast pork's welfare and the thermostat cannot even represent welfare, only temperature.

What, then, is welfare? When I'm concerned for my daughter's welfare as a result of her temperature, part of what I care about is alleviating her suffering.[32] She is in pain and I

[31] Dennett (1987)
[32] Bentham (1780/2007), Singer (1979)

would like to see this pain go away. And pain is a mental state. Is this true of welfare generally – does welfare consist in mental states?

Certainly, at least part of what constitutes welfare is mental states. Concern for making others' happy, or to help them feel safe, or to assuage their anxiety, this is concern for their welfare and it is directed at the mental states of happiness, feelings of safety, and anxiety.

But welfare also seems to include bodily properties in addition to mental states. In particular, it seems that an individual's health is an important component of their welfare.[33] Part of what I am concerned with when I am concerned about my daughter's temperature, for instance, is that she does not have an infection. On the face of it, this seems separate from my concern that she is not suffering, though I suspect that this is not the case.

As an infection may cause suffering, my concern that my daughter doesn't have an infection is to some degree constituted by my concern that she doesn't suffer. And insofar as an infection *doesn't* cause suffering – for instance, if the infection has rendered her unconscious and thereby unable to experience suffering – then it still prevents her from having positive experiences. In either case, ill health affects one's mental states and at least part of the reason why we care about it is because of these effects.

Given this, does it make sense to be concerned about health independently of concern about mental states? A farmer could care about the health of his crops without thinking that these crops have any mental states. But in this case, I'm inclined to think that his concern is derived from the lives that depend on the crops, and in particular, their mental lives. If he uses the crops for feed, then he is concerned to prevent the suffering of his livestock. If he sells the crops to support his family, then he is concerned with preventing his family from suffering in poverty. And if nobody depends on the health of his crops – if growing them is a mere hobby – then I'm inclined to think that his concern for their health is not a moral concern but the same kind of concern one might have about any hobby, in the same way that an ice skater might be concerned about her skates.

Given this, I'm inclined to think that concern for others' welfare, insofar as this is a *moral* concern, is concern for others' mental states. The implication, then, is that insofar as morality is concerned with others' welfare, moral agents must be able to have desires about others' mental states.

If it is the case that moral agents need to be capable of having desires about others' mental states, then this gives us a clear way forward for identifying moral agents: analyse the ability to have desires about others' mental states and identify what kinds of creatures have this ability.

Before we do this, though, there remains the question of whether one can be a moral agent *without* the ability to have desires about others' mental states. I have already suggested that concern for welfare is grounded in concern for mental states but morality is traditionally concerned with more than just welfare. Two other properties have traditionally been seen as legitimate objects of concern for morality: autonomy and fairness.[34] If it is possible for a moral agent to be concerned with one of these properties and not with welfare, and if concern for either of these properties is not grounded in concern for mental states, then

[33] Mill (1863/2002)
[34] Baron, Pettit, & Slote (1997), Parfit (2011), Cullity (2018)

this would imply that one could be a moral agent without the ability to have desires about others' mental states.

As I shall argue, concern for both autonomy and fairness require the ability to have desires about others' mental states. Thus, moral agents, understood as agents who have concern for either welfare, autonomy, or fairness, must have the ability to have desires about others' mental states.

Let's consider autonomy first. Autonomous individuals have certain *mental* abilities, most notably concerning deliberation and self-control. Violations of autonomy involve a disregard for these mental capacities,[35] such that the autonomous individual is treated as though they lack autonomy. For instance, failing to obtain informed consent for a medical procedure is a violation of the patient's autonomy because it ignores the patient's ability to deliberate about the procedure and come to their own decision.[36] By contrast, animals are not violated when they are treated without informed consent because they lack the relevant deliberative capacities.

If violation of autonomy involves a disregard for certain types of mental states, then it follows that respect for autonomy involves a regard for these same mental states. For a moral agent to respond to concerns about another individual's autonomy requires them to consider the relevant mental capacities so as not to bypass those capacities. For instance, a doctor seeking informed consent from a patient, desires (or ought to desire) that the patient can consider (that is, deliberate about) the procedure. It is possible for a doctor to seek informed consent without such a desire but this only indicates that he acts not out of a respect for autonomy but for some other reason, perhaps the prudential reason of retaining his medical license.

The upshot, to be clear, is that respect for others' autonomy requires an agent to have desires about others' mental states. These mental states may be different from the mental states relevant to welfare, but in both cases an individual incapable of responding to others' mental states cannot respond to the relevant concern, be it welfare or autonomy.

What about fairness? Can an agent respond to concerns about fairness without having desires about others' mental states? Given that fairness is about the distribution of benefits and burdens,[37] the best place to start is with the observation that for an agent to respond to concerns about fairness, they must recognise benefits and burdens.

Many benefits and burdens are directly related to welfare and autonomy. Improving another's welfare or granting them autonomy over their actions is a way of benefiting them. Reducing their welfare or violating their autonomy is a way of burdening them. Given that welfare and autonomy are benefits, and the lack of welfare and autonomy are burdens, an agent who failed to recognise welfare and autonomy would not be able to respond to concerns of fairness relating to the distribution of welfare and autonomy.

[35] The term 'autonomy' is an ambiguous one. Arpaly (2002, 2004) distinguishes several different concepts that have been called 'autonomy', of which *normative autonomy* is the concept that I am interested in here. Normative autonomy is what is violated when one dominates or manipulates others. She contrasts this with other concepts, including agent autonomy, as exemplified by Frankfurt's (1971) requirement for persons to have second-order volitions, and personal efficacy, by which she means the ability, lacked by young children, to take care of oneself. Throughout this thesis, I shall use the term 'autonomy' to refer to normative autonomy.

[36] Eyal (2019)

[37] Broome (1991), Rawls (1971/1999)

Insofar as fairness concerns the distribution of welfare and autonomy, a moral agent would need to have desires about others' mental states in order to respond to fairness.

This leaves open the possibility of benefits and burdens unrelated to welfare and autonomy. If such benefits and burdens exist, then perhaps it is possible for moral agents to respond to concerns about fairness even without the ability to have desires about others' mental states. Perhaps this is the case for the distribution of resources, such as food. While food is in most cases a means to improved welfare, one need not know this in order to care about the distribution of food.

Let's consider a case in which agents seem to be concerned with the fair distribution of food but whose concern does not rely on their having desires about others' mental states. The primatologist Frans de Waal and his colleagues have conducted an experiment with capuchin monkeys, which he takes to suggest that these monkeys have a sense of fairness.[38] In order for this experiment to constitute a genuine counterexample to the claim that concern for fairness requires the ability to have desires about others' mental states, I shall assume that these monkeys lack the ability to represent others' mental states.[39]

In this experiment, two monkeys each performed the same task, after which they were unequally rewarded with food. Both monkeys were within each other's sight as they performed the task and as they were rewarded, so they could see that the tasks were the same and the rewards differed.

The first monkey received a slice of cucumber after performing the task, whereas his companion received a grape, which is sweeter and thus more preferable to cucumbers. We are interested here in the behaviour of the first monkey, who upon seeing that the other monkey received a grape, rejected the slice of cucumber and presented his hand to receive a grape instead. In some trials, the monkey appeared to angrily throw the cucumber slice at the experimenter and scream and shake his cage. De Waal claims that the monkey rejected the cucumber slice because it was unfair to receive an unequal reward, and likened the sentiment expressed by the monkey to that of the Occupy protesters, who camped for several weeks in 2011 at Zuccotti Park near New York City's Wall Street to protest wealth inequality.[40]

But this does not strike me as the best interpretation of the experiment. Tellingly, only the first monkey rejects the reward. His companion is content to eat the grape, despite having witnessed the same inequality. A simpler explanation, then, is that both monkeys merely desire the grape, rather than a fair distribution of food. A desire for goods possessed by others is not on its own a desire for a fair distribution of goods.

So, what would a desire for a fair distribution of goods look like? What would be the *content* of such a desire? The monkeys have a desire for certain *goods*, but the desire for a fair distribution of goods also refers to the concepts of *fairness* and *distribution*. It strikes me that any desire for a fair distribution of goods must have content that refers to each of these concepts.

It might be argued that the monkeys desire a *particular* distribution of goods, namely the distribution in which the monkey doing the desiring possesses a grape. This seems to involve a very impoverished understanding of the concept of a distribution, though. By

[38] de Waal, *et al.* (2008)
[39] It does not matter for my present purposes whether capuchin monkeys can *actually* represent others' mental states. That said, in Chapter Four I discuss animals' ability to represent others' mental states and claim that the available evidence suggests that capuchin monkeys lack this ability.
[40] de Waal. (2011)

the same logic, one could claim that a thermostat desires the distribution of heat in which it is sufficiently warm to turn off the heating element. Given this, I don't think that the monkeys desire a *distribution* of goods, rather than merely desiring the goods themselves.

A desire for a particular distribution of goods seems to require reference to the beings to which the good may be distributed. If the first monkey desired that he have the grape and that the second monkey not have the grape, then this would be a desire for a particular distribution of resources.

It should be noted, though, that the reference to beings to which the goods may be distributed does not necessarily imply a reference to others' mental states. A farmer may desire a particular distribution of foxes on his farm: some in the wheat paddock to control the mice population and none in the chicken yard to prevent the chickens from being eaten. The farmer doesn't care about the mental states of the foxes, only that they perform the desired function in the desired place. The farmer could very well desire a particular distribution of rain for analogous reasons.

A desire for a *particular* distribution of resources does not require reference to others' mental states, but what about a desire for a *fair* distribution of resources. This turns on what is meant by 'fair' in this context. It strikes me that while fairness is related to equality, there must be something additional that distinguishes a desire for a fair distribution of goods from a merely equal distribution of goods. For instance, if a fair distribution of goods is merely an equal distribution of goods, regardless of the type of goods in question, then we would consider the distribution of *atoms* among human beings as fundamentally unfair. People who own more things will, in general, own more atoms, but few people care about the distribution of atoms over and above the distribution of the goods that are constituted by the atoms.

This suggests that for goods to be distributed fairly, as opposed to merely equally, they must be goods of a certain type. A good that nobody cares about is not the type of good that can be distributed fairly or unfairly. Given my discussion above of fairness applying to benefits and burdens, one might be inclined to think that this is because differences in the distribution of certain goods (such as atoms) do not differentially benefit or burden individuals, but I don't think this is quite right. After all, having more atoms means having more stuff, and having more stuff generally confers some benefit.

Rather, the defining feature of goods that can be distributed fairly seems to be that they are goods that agents *care about*. Unless one were to care about atoms, or to conceive of atoms as the kind of thing that should be cared about, one would have no reason to care about any particular *distribution* of atoms, such that some distributions were fair and others unfair. Given this, concern for fairness requires conceiving of the relevant good as being an target for the attitude of caring. Or, to use the language of representation: concern for fairness requires *mentally representing the relation between* caring and certain goods. Which is to say that concern for fairness requires mental representation of the attitude of caring, a mental state.

This gets us most of the way to the claim that concern for fairness requires desires about others' mental states, but there are two ways in which it still falls short. First, one can mentally represent a mental state, such as the attitude of caring, without having a *desire* about this mental state. And second, one can have a desire about mental states without these necessarily being *others'* mental states. I contend that concern for fairness does not fall short in either way.

Let's consider how concern for fairness involves *desires* about mental states. Recall that desires motivate agents to bring about a desired state of affairs by *representing* this desired

state of affairs. For instance, my desire for coffee is a mental representation of coffee (or, strictly speaking, the state of affairs in which I have coffee), which motivates me to get coffee. And my desire for fairness is a *mental representation of fairness* that motivates me to bring about a fair outcome. As I have claimed above, this mental representation of fairness includes reference to the attitude of caring. Thus, concern for fairness involves a desire about mental states.

But does it involve a desire about *others'* mental states? I think it does, but not in the straightforward way that concern for welfare or autonomy involves desires about others' mental states. In these cases, one desires certain things about *particular others'* mental states. If I am concerned for your welfare then I desire that *you* do not suffer. Concern for fairness is not always like this. In cases where one is concerned with the fair distribution of goods (as opposed to directly mental benefits and burdens, such as pleasure and suffering), and where one is concerned with fairness in general (as opposed to fairness *for* individuals, such as ensuring that you receive your fair share), there is no particular other for whom one has a desire about their mental states.

So how does concern for fairness involve desires about others' mental states in these cases? The answer, I think is that it involves desires about the mental states of a *generalised other*. If I desire a fair distribution of money, for instance, then the *content* of this desire includes the attitude of caring, because this desire is a desire for a fair distribution of money, where money is necessarily understood as something that people care about. Thus, the desire refers not to any particular person's attitude of caring, but to the attitude of caring in general, as applied to money. This is a desire about *others'* mental states, given that it is about a generalised attitude of caring, as opposed to *one's own* attitude. Thus, like desires for welfare and autonomy, the desire for fairness necessarily refers to the mental states of others, albeit in a generalised sense.

A Brief Note: Children and Animals

As mentioned earlier, Arpaly and I come to different conclusions regarding the presence of moral agency in children and animals. I believe these differences arise out of our differing views on which kinds of background conditions are necessary to be motivated by moral reasons.

Recall that Arpaly has claimed that children become moral agents around the time they learn the correct usage of the phrase "it's not fair".[41] I have argued, however, that the ability to respond to concerns of fairness is one of three ways in which one may be morally motivated, along with concerns of welfare and of autonomy. Although responding to any of these types of moral concern involve desires about others' mental states, there may be significant differences between the mental states involved in fairness than those involved in either welfare or autonomy, such that responding to concerns of fairness involves a level of cognitive complexity that is absent in children who can respond to other moral concerns. In Chapter Four, I develop this argument in detail. Specifically, I claim that responding to certain considerations of welfare involve cognitive capacities that are present in children as young as 18 months, whereas responding to concerns of fairness requires cognitive capacities that do not develop until around 3.5 years.

However, this is not all Arpaly has said on the topic of moral agency in children and animals. Her main argument against the moral agency of animals is that they lack the necessary concepts to be motivated by moral reasons. She gives the example of a young

[41] Arpaly (personal communication – 2019)

child who is upset that her dog has destroyed her favourite dinosaur toy. Arpaly imagines the child's parent explaining that the dog doesn't understand 'mine', 'favourite', or 'dinosaur'. Arpaly continues:

"Similarly, the dog's mind presumably cannot grasp – nor can it track, even in the way unsophisticated people can – such things as increasing utility, respecting persons, or even friendship. […] [E]ven if some protoversions of these notions exist in the animal's mind, these are not concepts that it can sophisticatedly apply to humans. Thus, even if this animal can act for reasons, to some extent, it cannot respond to *moral* reasons, even though it may sometimes come close."[42]

I agree with Arpaly that the dog likely lacks most of these concepts. Although the notion of a concept is difficult to pin down, for present purposes I am interested in the ability to discriminate between *types of things*. If I can discriminate between objects of a certain type and objects not of that type, then (for our present purposes) I have a concept of it. Thus, if the dog could discriminate between 'mine' and 'not mine', then it seems fair to say that it has some concept of ownership.[43]

That said, I disagree with Arpaly that possession of these specific concepts is necessary for moral motivation. As argued above, the relevant concepts are those related to welfare, autonomy, and fairness, where these are understood as mental states (such as suffering or pleasure) or as relationships between mental states and the world (such as deception as involving false beliefs).

Mark Rowlands has claimed that animals are capable of being motivated by compassion. He gives many examples, including an elephant who tries to help a dying relative to stand; a captive gorilla who rescued a young boy who had fallen into her enclosure, carrying the unconscious child to an access gate; and a golden retriever who saved a boy from a cougar at great risk to herself.[44] Insofar as these cases involved a desire about the others' welfare, it seems reasonable to agree with Rowlands that these animals were motivated by compassion.

Of course, there is a difference between the compassion of a human being who gives aid or comfort to another and the maternal instinct of, say, a bird feeding her chicks. One may think that the cases described by Rowlands are more like the latter than the former. However, as argued earlier, a relevant feature of praiseworthy compassion is that it is directed at another's mental state. Wanting to alleviate another's suffering is praiseworthy; wanting to feed one's chicks – in the absence of any desires about their hunger or comfort – is not.

This implies that the cases described by Rowlands would count as cases of moral motivation only if the respective animals were capable of having desires about others' mental states. As I shall argue in Chapter Four, this may well be the case for the elephant and the gorilla, but probably not the case for the dog.[45]

---

[42] Arpaly (2002), p. 146

[43] There is disagreement over whether animals have concepts at all. See, for instance, Davidson (1982) for the argument that they don't and Allen (1999) for the argument that they do. For the most part, I aim to avoid talk of concepts in this thesis, and instead talk about mental representation, which seems unambiguously present in many animals.

[44] Rowlands (2012)

[45] Although Rowlands (2012) argues (convincingly, I think) that animals can be motivated by moral reasons, he claims that they are not moral agents. This is because he favours an account of moral responsibility in which agents are only praiseworthy or blameworthy for their actions if they exert

Conclusion to Chapter One

I have argued that moral agency – in the minimal sense of an individual being an appropriate target of moral evaluations – requires the ability to form desires about the mental states of others.

To recap, I have argued, first, that there are several different ways one may respond to moral agents, but that the most basic is to evaluate them as either praiseworthy or blameworthy. I claimed that such evaluations depend first on evaluating their actions as right or wrong.

Second, and following Arpaly, I argued that agents merit praise for acting rightly insofar as they act from good will, and that they merit blame for acting wrongly insofar as they act from either ill will or insufficient good will. I noted a distinction between blameworthy and blameless acts of insufficient good will, and I argued that the difference lay in blameworthy agents being able to act out of good will.

Third, I argued over the next two sections that the ability to act out of good or ill will was an ability to act from desires with the right kind of representational content. I argued that this content is distinguished from non-moral content insofar as it picks out essentially moral properties. After examining three foundations of morality, welfare, autonomy, and fairness, I argued that concern for any of these foundations involves reference to others' mental states. Thus, I have argued that moral agency in the most inclusive sense requires the ability to have desires about others' mental states.

In the following chapter I extend this analysis of moral agency by claiming that this ability to have desires about others' mental states is not only necessary for moral agency in this inclusive sense but also sufficient for it.

the right kind of control over their actions (pp.88-93). As I shall argue in Chapter Two, I do not take this to be a necessary condition for the basic account of moral agency developed in this chapter.

CHAPTER TWO: UNNECESSARY CONDITIONS FOR BASIC MORAL AGENCY

In the previous chapter I introduced the concept of a moral agent – an individual that can be the appropriate target of moral evaluations, such as praise and blame. I argued that such agents must be able to respond to moral reasons, which I claimed were other-regarding.

This is a *quality of will* account of moral agency, as it justifies moral evaluations of agents in terms of their quality of will – the content of their desires. Agents with good will are appropriate targets of positive moral evaluation, such as praise, and agents with ill will or a lack of good will are appropriate targets of negative moral evaluations, such as blame.

This account owes a great deal to Nomy Arpaly's quality of will account and aims to extend it by offering a demarcation criterion between moral agents and other agents, such as infants, that are not appropriate targets of moral evaluation. I claimed that the ability to act out of good will and out of ill will consists in this ability to have desires about others' mental states, as this ability is necessary for responding to the three broad areas of moral concern: welfare, autonomy, and fairness.

In this chapter, I aim to make the case that this same ability – the ability to act out of good or ill will (as realised by the ability to have desires about others' mental states) – is not only a necessary condition for moral agency but is also a *sufficient* condition. As such, my main target in this chapter will be various alternative accounts of moral responsibility that propose additional conditions for moral responsibility.

A brief note is in order here regarding the shift above from discussing *moral agency* to discussing accounts of *moral responsibility*. As discussed in the previous chapter, the reason for this is that there is simply greater discussion of responsibility in the literature and that accounts of responsibility are essentially also accounts of agency, given the close connection between the conditions for moral responsibility and the conditions for moral agency. To recap, an account of moral responsibility spells out the *conditions* under which an agent is praiseworthy or blameworthy for performing an action, whereas an account of agency demarcates the *kinds of individuals* that can be appropriately held to be praiseworthy or blameworthy from those that cannot. The former can inform the latter insofar as the conditions for praiseworthiness and blameworthiness are properties of the agents themselves. As mentioned previously, moral agents are those individuals with the ability to meet the relevant responsibility conditions.

If an account of moral responsibility offers only global conditions, such as simple incompatibilism, then this cannot shed much light on questions of agency except to hold that their existence depends on the truth of indeterminism.[1]

Most contemporary compatibilist accounts, however, offer specific responsibility conditions that can be treated as conditions of agency.[2] Consider Arpaly's quality of will account (and my version of it), which offers responsibility conditions – agents can be praiseworthy or blameworthy on account of their acting out of good will, ill will, or a lack

[1] Of course, this is not to say that this is true of specific versions of incompatibilism. Robert Kane's libertarianism (in Fischer *et al.* (2007)), which includes the condition that responsible agents must have performed "self-forming actions", can distinguish between agents and non-agents on the basis of their ability to perform such actions. Such accounts can shed light on the demarcation criteria for moral agency precisely because they offer specific responsibility conditions (in addition to global responsibility conditions).

[2] McKenna & Coates (2021)

of good will; – thereby specifying an *ability* required of moral agents, the ability to act of good will or ill will.

Other accounts of moral responsibility can similarly be used to specify conditions of agency; moral agents are those individuals with the ability to meet the relevant responsibility conditions. In this chapter I consider several types of responsibility condition and argue that they are not necessary for moral responsibility, and therefore not necessary for moral agency. These are the *historical*, *epistemic*, *endorsement*, and *control conditions*. Other than quality of will accounts of moral responsibility, which do not always feature any of these conditions, almost all other compatibilist accounts of moral responsibility[3] tend to feature at least one of these conditions.

The Historical Condition

If I were to think of a theory of moral agency that *least* resembled my own, I would be hard-pressed to find something more divergent than Galen Strawson's *impossibilism*. While my account sets a very low bar for individuals to count as morally responsible agents – they need only be capable of having desires about others' mental states – Strawson sets his bar impossibly high. On Strawson's view, there are no circumstances under which anyone would be morally responsible for their actions.

As a general rule, I aim to keep my analysis focused on theories that share a certain amount of similarity with my own. I certainly don't expect to convince anyone already persuaded by Strawson's impossibilism of its falsity. Moreover, the broader debates in the literature, most notably the debate between compatibilists and incompatibilists, are well-worn and I don't expect to add anything new here. But because of its dissimilarity, impossibilism offers the starkest example of a condition that is shared by many compatibilist theories. I shall call this the *historical condition*, the claim that an agent's responsibility for their actions is mitigated by factors in their causal history that are outside of their control.

Given that moral agents are those individuals with the ability to meet the relevant responsibility conditions, if the historical condition is a genuine responsibility condition, then this implies that for one to be a moral agent, one must have a particular kind of causal history. Certain types of causal history, such as a history of abuse and neglect during one's formative years, may preclude one from being a moral agent.

In this section, I will consider this condition as it appears in Strawson's impossibilism, what I take to be the issues with this condition, and then I will discuss the condition as it appears in compatibilist theories that are much more similar to my own account of moral agency. I will claim that the same issues arise here and that we should therefore abandon the historical condition as a criterion for what it is to be a moral agent.

The *Basic Argument* is Strawson's argument for his impossibilism. We can express it as a simple *modus ponens*:

(1) If one is truly morally responsible for what one does, then one must be truly responsible for the way one is.
(2) But one cannot be truly responsible for the way one is.

[3] To be more accurate, one or more of these conditions form part of almost all accounts of *free will* or moral responsibility. Accounts of free will are important, however, because they are typically taken to specify, either in whole or in part, conditions for moral responsibility.

(3) Therefore, one cannot be truly morally responsible for what one does.

In defence of (2), Strawson points to various historical factors that prevent one from being responsible for the way one is:

"It is undeniable that one is the way one is, initially, as a result of heredity and early experience, and it is undeniable that these are things for which one cannot be held to be responsible (morally or otherwise). One cannot at any later stage hope to accede to true moral responsibility for the way one is by trying to change the way one already is as a result of heredity and previous experience. For both the particular way one is moved to try to change oneself, and the degree of success in one's attempt at change, will be determined by how one already is as a result of heredity and previous experience. […] It may be that some changes in the way one is are traceable not to heredity and experience but to the influence of indeterministic or random factors. But it is absurd to suppose that indeterministic or random factors, for which one is ex hypothesi in no way responsible, can in themselves contribute in any way to one's being truly responsible for how one is."[4]

I am inclined to grant, for the sake of argument, Strawson's defence of (2). Neither causal determinism nor indeterminism allow for the possibility of this deep sense of self-constitution.[5] That said, of course people can change who they are – for instance, by seeking an education – but they cannot change who they are *independently of who they already are*, which is the result of heredity, early experience, and possibly indeterministic factors, all of which are out of their control.

Strawson's defence of (1), however, marks an important point of difference between historical theories and pure quality of will theories. The link between responsibility for one's self-constitution and responsibility for one's actions relies on the justification of punishment. If one is not *ultimately responsible* for one's actions, that is, responsible in the deep sense linked with responsibility over one's self-constitution, then punishment cannot be justified. Strawson illustrates this relationship in two passages:

"As I understand it, true moral responsibility is responsibility of such a kind [the kind that is both impossible and widely believed in] that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of use with (eternal) torment in hell and reward other with (eternal) bliss in heaven. […] The story of heaven and hell is useful simply because it illustrates, in a peculiarly vivid way, the *kind* of absolute or ultimate accountability or responsibility that many have supposed themselves to have, and that many do still suppose themselves to have. It very clearly expresses its scope and force."[6]

"We are what we are, and we cannot be thought to have made ourselves *in such a way* that we can 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."[7]

Strawson makes an interesting shift in these two passages. In the first, he motivates belief in *ultimate moral responsibility* as the only possible justification for the eternal torment of hell, what we might call *ultimate punishment*. But in the second passage, he claims that this kind of ultimate moral responsibility is required to justify *any punishment*. It's not at all clear to me that this is the case. Perhaps there is a limited sense of responsibility that justifies a limited, earthly punishment.

[4] Strawson (1994)
[5] See also Levy (2011)
[6] Strawson (1994)
[7] *Ibid.*

Neil Levy, another responsibility sceptic, bridges this gap by arguing that the sense in which it doesn't make sense to punish individuals, or to even hold them responsible, is because it would be *unfair* to do so:

"[V]ery often we cannot ignore questions of history. Agents acquire their responsibility-relevant characteristics – their characters, their resources of self-control, their values and beliefs – as a consequence of their socialization. The resources they utilize when they act are themselves socially distributed. From the agent's point of view, these resources are acquired luckily; as the product of present, or, more usually, constitutive luck. But from the society's point of view they are not merely lucky: they are predictable consequences of social choices. Agents with false moral views, views they take to justify actions that are actually immoral, have often been deprived of the opportunity to acquire more accurate beliefs; agents who lack the resources of self-control have had the bad luck to occupy social roles from which self-control cannot easily be acquired."[8]

This approach – of requiring a historical condition for moral responsibility to make sense of punishment – is not limited to incompatibilists. John Martin Fischer, a prominent compatibilist,[9] also holds this position. Fischer's views are particularly interesting because, like Arpaly, he claims that the key feature of moral responsibility is agents' ability to respond to reasons. On his view, it is this reasons-responsiveness that makes punishment appropriate for moral agents and not for creatures who are unable to respond to reasons:

"[Punishment] affects the desirability of performing a certain action. That is, punishment involves reacting to persons in ways to which the mechanisms on which they act are sensitive. My suggestion is that punishment is appropriate only for a creature who acts on a mechanism "keyed to" the kind of incentives punishment provides. My point here is not that the justification of punishment is "consequentialist" – that it alters behavior. (Of course, this kind of justification does not in itself distinguish punishment from aversive conditioning.) … My justification is nonconsequentialist and "direct": punishment is an appropriate reaction to the actual operation of reasons-responsive mechanisms. When it is justified, punishment involves a kind of "match" between the mechanism that produces behavior and the response to that behavior."[10]

This notion of a mechanism that is 'keyed to' the incentives of punishment explains why agents must be responsive to reasons in order for their punishment to be justified. For Fischer, punishment is a *provider of reasons*. Only if an agent is responsive to the reasons provided by the threat of punishment (by possessing a mechanism 'keyed to' its incentives), can punishment be justified.

Fischer explicitly links the reasons-responsiveness element of his account not just to the justification of punishment, but also to the claim that accounts of moral responsibility must take into consideration an agents' history. In doing so, he distinguishes his account from purely structural accounts such as those of Harry Frankfurt and Gary Watson:[11]

"I wish to contrast my approach to moral responsibility with a class of theories that might be called "mesh" theories of responsibility. My approach is a historical theory. Consider

[8] Levy (2011), p.195

[9] Strictly speaking, Fischer considers his position to be *semicompatibilism*, emphasising that it has both compatibilist and incompatibilist elements. However, the sense in which it is incompatibilist is that it accepts Peter van Inwagen's *consequence argument*, which holds that for any action, given the truth of determinism, we could not have acted otherwise. Fischer's account is compatibilist in the sense that it regards moral responsibility as compatible with determinism (in Fischer *et al* (2007))

[10] Fischer (1997), p.80

[11] See Frankfurt (1971) and Watson (1975)

first a "hierarchical" model of moral responsibility. In this model, a person is morally responsible for an action insofar as there is a mesh between a higher order preference and the first-order preference that actually moves him to action. […] The problem with such hierarchical "mesh" theories, no matter how they are refined, is that the selected mesh can be produced via responsibility-undermining mechanisms. After all, a demonic neurophysiologist can induce the conformity between the various mental elements via a sort of direct electronic stimulation that is not reasons-responsive. I believe that the problem with the hierarchical mesh theories is that they are purely structural and ahistorical. It matters what kind of process issues in an action. Specifically, the mechanism issuing in the action must be reasons-responsive. […] The mesh between the elements of different preference systems may be induced by electronic stimulation, hypnosis, brainwashing, and so on. Moral responsibility is a *historical* phenomenon; it is a matter of the kind of mechanism that issues in action."[12]

At this point, one may notice that Arpaly and Fischer seem to be referring to slightly different things when they speak of 'reasons-responsiveness'. Fischer emphasises the causal history of actions, as issuing from a mechanism that is responsive to reasons, whereas Arpaly is concerned with an identity relation between agents' reasons for action and reasons for which their action is good or bad. For Fischer, reasons-responsiveness is essentially historical, while for Arpaly, it is not.

This point is important because Fischer's manipulation argument against mesh theories generalises to any ahistorical account, including accounts in which reasons-responsiveness is ahistorical. For instance, an evil neurosurgeon could in principle induce in an agent the desire to respond to egoistic reasons more readily than moral reasons. I shall consider this possibility in my discussion of the control condition later in this chapter.

But for Fischer, whose account of reasons-responsiveness is historical, there is a clear link between this historical condition and punishment. In short, if an agent acts wrongly, it is only appropriate to punish them if they could have responded to the relevant reasons, and they could only have *really* responded to the relevant reasons if they were not a victim of manipulation at the time. The historical condition, then, is needed to ensure that agents are not punished for actions resulting from manipulation.

When it is spelled out like this, the problem becomes apparent. Punishment is not the same thing as blame, so even if punishment requires the historical condition, this does not imply that blame also does.

But to see how blame differs from punishment in this respect, consider again the types of blame outlined in the previous chapter:

1. The mere evaluation of the agent's badness
2. An emotional response, such as indignation
3. The expression of an emotional response, such as a look of disapproval
4. An utterance, such as "Brutus is to blame for the death of Caesar"
5. Social censure arising as the result of such expressions and utterances, either intentionally or inadvertently
6. Punishment proper, wherein the blameworthy individual is seen as deserving rebuke, and punished accordingly, on the basis of his blameworthiness

Now consider Levy's claim that blame is often unfair on the grounds that it confers benefits and burdens to individuals who had no control over the historical conditions that led to their blameworthiness. This strikes me as true of senses 3-6 above, in which this

[12] Fischer (1997), p.79

blame is expressed in ways visible or potentially visible to the person being blamed. But it seems senses 1 and 2, which are characterised by evaluations and attitudes entirely internal to the person doing the blaming, confer no such benefits or burdens.

Given that I am primarily concerned with blame in the first sense, blame-as-evaluation, it is worth asking what makes blame in *this* sense unfair. Arpaly has a response to this question:

"The primary sense in which I can be fair or unfair in blaming someone is the sense in which believing that Ron is an idiot might be fair if Ron is an idiot and unfair if Ron is not. The primary sense in which I can be fair or unfair in punishing someone is the sense in which my calling Ron an idiot might be fair if he has just called me a moron and unfair if he has never been rude to me."[13]

It would be unfair to punish someone for something they haven't done, just as it would be unfair to call someone an idiot if they have never been rude. On this point, Arpaly and Levy, seem to agree. And if 'something they haven't done' is taken to include actions over which one has no control, then it seems reasonable to say that agents should not be punished for these actions either.

But, as Arpaly points out, the grounds for fairly blaming someone differ from the grounds for fairly punishing someone. Blame-as-evaluation is analogous to belief, insofar as both can be *warranted*, depending on whether the target meets the criteria of the evaluation. Thus, I am justified in believing someone is an idiot if they are an idiot, and I am justified in blaming someone if they are blameworthy. As discussed in the previous chapter, an agent's blameworthiness depends on their quality of will; it does not depend on the historical conditions that gave rise to their quality of will.

The Epistemic Condition

Susan Wolf is another compatibilist who, like Fischer and Arpaly, takes reasons-responsiveness to be central to her account of moral responsibility. Unlike Fischer, Wolf's account of moral responsibility does not seem to be motivated by an interest in justifying punishment. In fact, one of her criticisms of early compatibilist accounts of moral responsibility, in which punishment is justified on consequentialist grounds, is that such theories are improperly motivated by an interest in justifying punishment:

"[Such an account][14], insofar as it is offered as a solution to the problem of responsibility, is naïve and simplistic because it fails to recognize that the concept of responsibility is connected to the practices of reward and punishment only by way of and in the company of this concept's connection to subtler, less overt practices that involve attitudes, for example, of admiration and indignation, and judgments of agents' deserts."[15]

---

[13] Arpaly (2006), p.9

[14] Here, Wolf is specifically referring to the account of Moritz Schlick (1963), whose account of moral responsibility is fairly typical of compatibilist accounts of the early- to mid-twentieth century (similar accounts include those of Hobart (1934) and Smart (in Smart & Williams (1973)). These accounts are generally characterised by a commitment to a consequentialist justification of punishment, such that agents ought only to be punished for an act of wrongdoing if doing so is likely to prevent their future wrongdoing. Given this, agents for whom the threat of punishment is not an effective deterrent, such as sufferers of certain compulsions or mental illnesses, ought not to be punished.

[15] Wolf (1990), p. 19

This is interesting because Wolf's best-known example of a non-blameworthy individual, an example that has since become a staple of the literature on moral responsibility, involves someone whose purported non-blameworthiness derives from the historical conditions of his upbringing:

"JoJo is the favourite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not *coerced* to do these things, he acts according to his own desires. Moreover, these are desires he wholly *wants* to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. In light of JoJo's heritage and upbringing – both of which he is powerless to control – it is dubious at best that he should be regarded as responsible for what he does."[16]

Given that Wolf's interest in developing a theory of moral responsibility is not in explaining the conditions under which agents deserve punishment, it is clear that her belief in JoJo's non-blameworthiness is not due to a confusion between the conditions for blame and the conditions for punishment. Why, then, does she take JoJo to be non-blameworthy?

For Wolf, JoJo's ignorance of right and wrong, or rather his inability *not* to be ignorant of right and wrong, given his upbringing, is sufficient to excuse him from blame. Like Arpaly, Wolf claims that praiseworthiness consists in the agent doing the right thing for the right reasons. It is therefore interesting that Wolf and Arpaly derive from this claim different conclusions about the blameworthiness of agents who lack knowledge of right and wrong.

Wolf argues that doing the right thing for the right reasons – acting out of good will, in Arpaly's language – requires both knowing what is good and converting this knowledge into action. Since JoJo's upbringing has made it impossible for him to know what is good, he cannot act out of good will and thus is not blameworthy for his wrongdoing.[17]

If the epistemic condition is a genuine responsibility condition, then this implies that for one to be a moral agent, one must have a particular kind of knowledge or the ability to acquire this knowledge. Certain epistemic defects, such as JoJo's inability to know what is right or wrong, may preclude one from being a moral agent.

But it is not necessary to *know* what is good in order to *do* what is good. More to the point, it is not necessary to know what is good in order to do the right thing for the right reasons. Arpaly illustrates this point with the case of Huckleberry Finn, who does the right thing for the right reasons without realising that he is doing so.

"Huckleberry Finn befriends Jim, a slave, and helps him escape from slavery. While Huckleberry and Jim are together on a raft used in the escape, Huckleberry is plagued by what he calls "conscience." He believes, as everyone in his society "knows," that helping a slave escape amounts to stealing, and stealing is wrong. […] [W]hen the opportunity comes to turn Jim in and Huckleberry experiences a strong reluctance to do so, his reluctance is to a large extent the result of the fact that he has come to see Jim as a person,

[16] Wolf (1987)
[17] Wolf (1990), p. 88

even if his conscious mind has not yet come to reflective awareness of this perceptual shift. To the extent that Huckleberry is reluctant to turn Jim in because of Jim's personhood, he *is* acting for morally significant reasons. This is so even though he does not *know or believe* that these are the right reasons. The belief that what he does is moral need not even appear in Huckleberry's unconscious."[18]

This strikes me as a plausible account of Huckleberry's moral psychology, and I think Arpaly is correct to claim on the basis of this example that one can do the right thing for the right reasons without knowing that one is doing so, and that one is praiseworthy for such actions.

This is all well and good for praiseworthy agents, and when it comes to Huckleberry Finn, even Wolf seems open to the possibility of his praiseworthiness.[19] What seems to elicit stronger intuitions, however, are cases in which agents act *wrongly* without knowledge of the wrongness of their actions.[20] As mentioned above, Wolf claims that the blameworthiness of such agents depends on whether they could have done the right thing for the right reasons, which further depends on whether they had the *ability* to know what is right.

So, is Wolf correct here? If an agent acts wrongly but could not have known any better at the time of the act, is this agent blameworthy? On my account, it depends on the grounds for blameworthiness, which differ for acts of ill will and for acts of insufficient good will. Recall that agents are blameworthy for acts of ill will because such agents *intend to act wrongly*, whether or not they see themselves as doing so. Agents are blameworthy for acts of insufficient good will *only if they could have done better*.

Thus, I agree with Wolf in the case in which an agent acts wrongly out of insufficient good will but could not have known better. On Wolf's analysis, such an agent could not have known better, and therefore could not have done better, and therefore is not blameworthy.

My agreement with Wolf in this specific case should not, however, be taken as endorsement of the epistemic condition generally. Blameworthiness for acts of insufficient good will is necessarily more complex than blameworthiness for acts of ill will, because the former presupposes the ability to act out of good will, whereas the latter does not.

When we consider agents who act wrongly out of ill will but who could not have known better – agents such as JoJo – my disagreement with Wolf becomes clearer. On Wolf's account, JoJo's inability to know better exempts him from blameworthiness, whereas on my account, JoJo's blameworthiness does not derive from his ability to know, or to do, better, but from his intending to do the wrong thing.

My contention, then, is that the epistemic condition is only a condition of blameworthiness for acts of insufficient good will, because blameworthiness in such cases presuppose an ability to do better. The epistemic condition is not a condition for blameworthiness for acts of ill will, nor is it a condition for praiseworthiness for acts of good will.

---

[18] Arpaly (2002), p. 75-77

[19] More accurately, Wolf seems *agnostic* on this point, as she suggests in passing by noting the complexity of such cases. That said, in the relevant passage (1990, p. 143), she does not make any definitive claims about the praiseworthiness of such agents, nor does she analyse these cases in detail.

[20] The relative abundance of thought experiments involving blameworthiness and wrongdoing in the responsibility literature, compared with those involving praiseworthiness and rightdoing, suggests that intuitions about the former are in general stronger than those about the latter.

The Endorsement Condition

Turning back again to Fischer's account of moral responsibility, recall that he contrasted his account with so-called "mesh theories", in which moral agents were distinguished from lesser agents by the structure of their mental states. Such theories include Harry Frankfurt's "second order volition" account, in which agents are characterised by the possession of desires about other desires, and which are effective in producing action.[21] My account, in which agents are also characterised by their ability to have desires with a particular content (others' mental states) and which are effective in producing action, is also such a mesh theory.

Fischer's criticism of mesh theories was that by failing to take into account an agent's history, these theories were vulnerable to responsibility-undermining manipulations. I claimed in that section that Fischer's account is similarly vulnerable to such manipulations. Fischer himself recognises this possibility. As a result, he offers an additional condition for blameworthiness. Rather than relying solely on the causal history of agents' desires and actions, Fischer also emphasises the agent's endorsement of her own actions:

"But one could exhibit the right sort of reasons-responsiveness as a result (say) of clandestine, unconsented-to electronic stimulation of the brain (or hypnosis, brainwashing, and so forth). So [the appropriate kind of reasons-responsiveness] is necessary but not sufficient for moral responsibility. I contend that there are two elements of guidance control: reasons-sensitivity of the appropriate sort and mechanism ownership. That is, the mechanism that issues in the behavior must (in an appropriate sense) be the *agent's own mechanism*. […] My co-author, Mark Ravizza, and I argue for a subjective approach to mechanism ownership. On this approach, one's mechanism becomes one's own in virtue of *seeing oneself in a certain way*. […] In our view, one becomes morally responsible in part at least by taking responsibility; he makes his mechanism his own by taking responsibility for acting from that kind of mechanism. In a sense, then, one acquires control by *taking control*."[22]

This concept of *endorsing* one's actions, along with similar concepts, such as *identification* with one's actions, is a common response to the issue of the problem of desires and actions from which agents feel alienated.[23] Paradigmatic cases of alienation include hypnosis and brainwashing, as well as the desires experienced by sufferers of compulsive disorders, such as kleptomania.

If the endorsement condition is a genuine responsibility condition, then this implies that for one to be a moral agent, one must take a particular attitude toward one's actions. The inability to take such attitudes may preclude one from being a moral agent.

For Fischer, one can only become responsible for one's actions by taking responsibility for them. While it may be magnanimous for an agent to take responsibility for the "alien" actions arising from brainwashing, hypnosis, or compulsion, he is not under any obligation

[21] Frankfurt (1971). Frankfurt doesn't himself specify these agents as *moral* agents, though Fischer seems to treat them as such for the purpose of his discussion, by distinguishing his own account of moral responsibility from Frankfurt's account of agency.
[22] Fischer (2004), p. 18; emphasis in original
[23] Harry Frankfurt, for instance, has written extensively on the concept of identification. See Frankfurt (1977, 1987)

to do so.[24] And so, for Fischer, alienation from one's actions would typically absolve one of responsibility.

I think this view is mistaken. In short, the mere fact that an action is *experienced* as alien does not make it so. There is a very real sense in which one's 'alien' actions are still one's own: they are performed by one's own body and are caused by one's own desires. Arpaly offers many examples to illustrate this point,[25] but two shall suffice to make it clear:

"If a Victorian lady experiences her sexual desires as alien, intrusive, "not truly her own," our natural reaction is to tell her she is wrong, that these desires are in fact her own, and that only the false, asexual self-image that she acquired with her upbringing makes her experience them as threatening to her integrity as a person." [26]

"As a rule, people seem to be able to feel alienated from many things – that is, to experience these things as at once belonging to them and alien. For example, a woman who used to be thinner in her youth may stand in front of the mirror and experience her fat thighs as "not really her own," as if someone has latched them onto her. It only means that the fatness of her thighs conflicts with her visceral image of herself. Similarly, I see no particular reason to believe that a desire from which one feels alienated is in any sense "less one's own" – all we know is that the desire conflicts with the person's visceral self-image, which may be accurate or, as in the case of the woman, misguided." [27]

Arpaly takes these examples to show that there are many things, including actions, which are experienced as both our own but also as alien, and that this experience is insufficient to exempt one from blameworthiness if the 'alien' action is performed out of ill will or insufficient good will. I think Arpaly is correct here. And this is why I think Fischer is mistaken in suggesting that an agent who does not take responsibility for his actions – even his 'alien' actions – is not blameworthy in virtue of this.

Although I think that the concepts of identification and alienation are not a useful addition to our account of moral responsibility, one still may have the intuition that victims of manipulation or compulsion are nonetheless not blameworthy for their wrongdoing.[28] The worry here may not be that actions caused by manipulation or compulsion are "not one's own" in the sense of being experienced as alien, but that such actions are *outside of the agent's control*, and therefore, that an account of responsibility must include a control condition to properly exempt victims of manipulation and compulsion from blame. It is to this concept of a control condition that I now turn.

The Control Condition

Now let's turn to cases of action outside the agent's control. The paradigmatic cases of such action tend to fall into two broad classes. The first are actions that are under *someone else's* control, such as actions arising from hypnosis and brainwashing. Let's call cases of this sort *manipulation cases*.[29] The second type of uncontrolled actions are those that issue solely from the agent but are, in a sense, not *of* the agent. The paradigmatic case here is that of compulsion, including compulsive disorders such as kleptomania, but other types

[24] Fischer (2004)
[25] See Arpaly (2002), pp. 123-131 for a discussion of these cases.
[26] *Ibid.*, p. 123
[27] *Ibid.*, pp. 130-131
[28] This intuition is widespread. In the following section, I will discuss *manipulation cases*, which have frequently been used to undermine accounts of moral responsibility.
[29] See, for instance, Mele (1995), Rosen (2002).

of behaviour may also fall under this description, such as the involuntary tics experienced by sufferers of Tourette's Syndrome. Let's call these *uncontrolled action cases*.[30]

If the control condition is a genuine responsibility condition, then this implies that for one to be a moral agent, one must have a particular kind of control over one's actions. Certain deficiencies in one's control, as is the case in manipulation cases and uncontrolled action cases, may preclude one from being a moral agent.

Let's consider manipulation cases first. In what follows, I will assume that such cases are not simply reducible to cases of uncontrolled action. That is to say, I will assume that manipulated agents still retain their rational faculties such that any responsibility-mitigating factors arising from their manipulation are not explainable in the same terms that might explain diminished responsibility in the case of uncontrolled action, such as compulsion. In short, I will assume that manipulators don't turn unwitting agents into kleptomaniacs, but into wholehearted thieves.

Cases of manipulation in the literature typically involve the manipulator inculcating desires in the targeted individual by one of a variety of nonrational means. These range from Frankfurt-style cases, in which a nefarious neurosurgeon implants a device that can produce an effective desire in the victim at the press of button, to Wolf's JoJo case, in which the affected individual is slowly and unintentionally brainwashed during his childhood.

We have already seen evidence that manipulation examples can conceivably be devised as counterexamples to almost any particular non-sceptical account of free will or moral responsibility, or as a counterexample to such non-sceptical accounts in general. For instance, if one wished to dispute Frankfurt's second-order volition account of free will, in which free will is constituted by the presence of a desire which is about another desire (a second-order desire) and which is effective in producing action (a volition), then the critic could offer the example of an individual who was inculcated with an errant second-order volition. One could use the same method to criticise almost any account of free will or moral responsibility, by taking the relevant criterion for responsibility or free will, and then offering the example of an individual who came to acquire this criterion though manipulation.[31] The effectiveness of this strategy depends on our intuition that manipulated individuals do in fact lack free will or moral responsibility.[32] My contention is that these cases are either misdirected or not as intuitively compelling as they at first seem.

Before I go on, I must raise a brief point about manipulation cases in the philosophy literature, specifically those involving brainwashing. As is often the case when philosophers discuss psychology, there is debate over whether brainwashing cases, as described by philosophers, reflect the reality, as described by those, such as psychologists, with empirical expertise in the phenomenon in question.[33] For my purposes, I'm not concerned with *actual* cases of brainwashing, but rather the cases as presented in the literature, since it is these cases that aim to elicit sceptical intuitions about moral responsibility and thereby require a response.

---

[30] See, for instance, Kennett (2001), Mele (2012).
[31] Stump (1996) has offered a manipulation argument in response to Frankfurt's (1971) account of moral responsibility, while Pereboom (in Fischer, *et al.* 2007) has offered a manipulation argument in response to compatibilist accounts in general.
[32] See, for instance, Vargas (2013).
[33] See, for instance, Black & Tweedale (2002).

Recall that the psychological structures taken to ground moral responsibility in so-called mesh theories of responsibility, including my own, can be produced in way that appear to undermine responsibility, such as by direct neural manipulation. An agent's motivation (her reason for action) could be produced by direct neural manipulation yet still bear the appropriate identity relationship with the right-making features of her act. That is, our Frankfurtian neurosurgeon could implant in our agent a desire to commit a specific wrong and if the agent were to act on this desire then they would be blameworthy on a quality of will account such as mine or that of Arpaly. Arpaly's response to manipulation arguments is not to propose additional conditions necessary for moral responsibility, but instead to bites this bullet. Although manipulation cases may appear to undermine responsibility or mitigate blameworthiness, this is in fact false:

"[C]onsider the case of Patty Hearst – perhaps as close as reality gets to [Frankfurt-style interventions]. Brainwashed by her captors, Hearst joined their terrorist organisation and was eventually convicted for her crimes despite the fact of her nonrational change in motivations. Note that it matters very little to our judgment if she has indeed been brainwashed deliberately or if she just converted, irrationally, due to the duress she was under (the "Stockholm Syndrome"). In either case, a drastic change in her belief-desire set happened irrationally and rather quickly, and in either case the person who stood before the court seems to have been a wholehearted terrorist who was blameworthy for her actions, not an innocent woman acting under great duress. Stress may cause some people to act out of character, but it may also truly *change their characters*, and this is what seems to have happened in the case of Hearst."[34]

I believe this to be the right response.[35] The mere fact that Hearst's desires and actions (or those of the victims of Frankfurt-style neural interventions) were ultimately derived from another individual has no bearing on her blameworthiness. There are several reasons for thinking otherwise, but these strike me as mistaken.

One reason for thinking that manipulated agents are blameless for their manipulated actions is that these actions derive ultimately from an *external source*.[36] However, in my discussion of the work of Galen Strawson, I argued that while the source of one's actions may be relevant in determining whether one deserves punishment, it tells us nothing about whether the agent is a bad person. In particular, it tells us nothing about whether his actions express the quality of his will.

A second reason for thinking that manipulated agents are blameless for their manipulated actions is that these actions are not *experienced* as their own. However, in my discussion of the work of Fischer, I argued that alienation from one's actions shows only that one's actions are in tension with one's self-image, and that this tension has no bearing on one's blameworthiness. Again, it tells us nothing about whether one's actions express one's quality of will.

A third reason, related to the first, is that certain *types* of action deriving from external sources exempt agents from blameworthiness. Specifically, actions deriving from *other individuals* undermine the blameworthiness of the manipulated individual. I think that this is also mistaken because there are no relevant differences between the "manipulations" of external sources that are not other agents, such as genetics and upbringing, and the

[34] *Ibid.*, p. 166

[35] This response, sometimes known as *hard compatibilism*, is endorsed by several compatibilists, including Arpaly (2005), McKenna (2005), and Russell (2010).

[36] Such arguments are typically offered by incompatibilists, including Chisholm (1964), Clarke (1993), and Kane (in Fischer *et al.* 2007), but compatibilists sometimes offer such arguments as well. See, for instance, Fischer (in Fischer, *et al.* (2007)).

manipulations of other agents. Derk Pereboom offers a good argument for this claim, comparing four different cases of manipulated and "manipulated" behaviour to show that there are no relevant differences between them that bear on the responsibility of the individuals in each case.

It must be noted, however, that Pereboom is a sceptic about moral responsibility, and uses the similarity between the four cases to argue from the intuition that manipulated agents are *not* blameworthy to the claim that causally determined agents are not blameworthy. It is worth discussing this argument because it strikes me that we can use it to infer a contrary claim: manipulated agents *are* in fact blameworthy.

Pereboom begins with the case of Plum, whose brain is remotely manipulated by a team of button-pushing neuroscientists, such that he desires to kill White. Pereboom stipulates that this intervention does not inculcate an irresistible desire and leaves his rational faculties intact. Plum's psychology is such that it conforms to the responsibility conditions of the major theories of moral responsibility: For our purposes, Plum is responsive to the relevant moral reasons, but his psychology is manipulated in such a way that he is more responsive to the egoistic reasons in favour of killing White. Surely, Pereboom assures us, Plum cannot be blameworthy for killing White in this situation.

We then move to the second case, in which Plum is not manipulated from moment-to-moment, but is instead programmed at birth to weigh reasons in such a way that he will act so as to kill White. Surely, Pereboom suggests, if Plum is not blameworthy in the first case then nor is he blameworthy here, as the only difference between the two cases is when the manipulation occurred, and this is not a morally relevant difference.

In the third case, Plum's behaviour is not the result of direct manipulation, but of his experiences during early childhood. Plum, like JoJo, is raised in such a way that he finds killing White more appealing than the alternative. Again, Pereboom claims that Plum is not blameworthy, as there is no obvious responsibility-relevant characteristic that he possesses here but not in Case 2. Pereboom observes that in both Cases 2 and 3, Plum meets the responsibility criteria for the major compatibilist theories of moral responsibility, so if a compatibilist were to exempt Plum from blame in Case 2, then he must do the same in Case 3.

Finally, Case 4 describes a situation in which Plum is causally determined to kill White, not because of any intervention by other agents, but purely because he is a physical being in a universe where the behaviour of physical systems is causally determined. Given that his behaviour is no less determined in this case than in any of the previous three, Pereboom claims that even in this case, Plum is not blameworthy for the death of White.[37]

I agree with Pereboom that there are no responsibility-relevant differences between any of the four cases. If Plum is not blameworthy for killing White in Case 1, then he is not blameworthy in any of the other cases either. Conversely, if Plum is blameworthy in any of the four cases then he is blameworthy in all of them.

Where I disagree with Pereboom is in his initial assessment of Plum's blameworthiness in Case 1. If I am correct about the general conditions for blameworthiness as outlined in the previous chapter, then Plum *is* blameworthy for White's death, regardless of how he came to have the desires that he has. But the spectre of manipulation is persuasive, and it is tempting to think that if someone else is blameworthy for Plum's actions – as the neuroscientists surely are – then this lessens Plum's blameworthiness. It does not. Although there is a very real sense in which the neuroscientists *made* Plum do it, *Plum* still

[37] Pereboom (in Fischer, *et al.* (2007)), pp. 93-98

did it. His rational faculties were intact at the time of his action, which was caused by Plum's desire for White's death.

But I understand that my intuitions about this case are not widely shared. If, as Pereboom intends, your intuitions about Plum in Cases 1 and 4 pull you in opposite directions, of judging him to be blameless in Case 1 and blameworthy in Case 4, then I encourage you to run the cases in reverse. If Pereboom is correct that there are no responsibility-relevant differences between the four cases, then we can just as easily use the cases to derive Plum's blameworthiness in Case 1 from his blameworthiness in Case 4.

Of course, this is not a novel insight,[38] and Pereboom has offered a response. He claims that the sequence is intended to make the deterministic nature of the causes of Plum's behaviour salient, and that to begin with the case in which this were not salient would beg the question against the incompatibilist.[39]

In response, I first note that Pereboom and I have different targets in our use of the four-case argument. He wishes to make determinism salient by beginning with an uncontroversial case of determined behaviour, and proceeding to progressively less obvious cases of determined behaviour to claim that our intuitions about Plum's blamelessness in cases of manipulation apply to the general case of determinism, even when he is not manipulated by other agents. I wish to make *reasons-responsiveness* salient by beginning with an uncontroversial case of reasons-responsive behaviour, and proceeding to progressively less obvious cases of reasons-responsive behaviour to claim that our intuitions about Plum's blameworthiness in cases where he is responsive to reasons apply in all cases where he is responsive to the relevant reasons, even when his behaviour is manipulated by other agents. Given my use of the four-case argument in making salient *reason-responsive behaviour*, it strikes me as question-begging against *this* to run the cases forward, rather than in reverse.

Moreover, it strikes me that it makes *more* sense to run the cases backwards rather than forwards. Both methods rely on our intuitions about the initial case, which is then generalised to the other three cases. While case (1) seems to elicit stronger intuitions than case (4) – the blamelessness of victims of direct neural intervention seems more *salient* than the blameworthiness of ordinary wrongdoers – it strikes me that the intuitions elicited by case (4) are more *reliable* than those elicited by case (1).[40] We all have experience of blaming and being blamed under ordinary circumstances but cases of direct neural intervention are rare and our intuitions regarding the blameworthiness of such individuals tend to be shaped by philosophers' assertions rather than our direct experience. Even the case of JoJo, taken here as an example of case (2), elicits more uncertainty in our intuitions than the ordinary case. It strikes me, then, that the mere fact of agents' manipulation is not enough to exempt them from blame.

In the arguments above, I have assumed that manipulated individuals still have their rational faculties intact. I have assumed, for instance, that Frankfurt-style neural manipulators have left their victims *able* to express good and ill will, but that that the *content* of this will has changed. In other words, I've assumed that agents retain their ability to respond to reasons but, as a result of the manipulation, they respond to different reasons than they otherwise would have. The Frankfurt-style neurosurgeon has not changed these agents into automata but rather has changed them into *bad people*. Thus, while they are not

[38] For instance, Pereboom quotes Michael McKenna (2005) as making the same suggestion.
[39] Pereboom (in Fischer *et al.* 2007), p.100
[40] See Vargas (2013).

responsible for being the way they are, we are still justified in believing them to be worthy of a poor moral evaluation, in believing them to be blameworthy.

But what of the agent who, as a result of manipulation, does become an automaton? Or, less drastically, one who acts wrongly as a result of an implanted compulsion? I contend that, just as reasons-responsive victims of manipulation are no more or less blameworthy than other reasons-responsive agents, compelled victims of manipulation are no more or less blameworthy than agents who suffer from internal compulsions.

The upshot of this is that we can treat compelled agents as a single class, regardless of whether the compulsion is internal or external in origin. Given this, we can now turn our attention away from manipulation cases and toward cases of compulsion generally, as well as other cases of seemingly uncontrolled action that have their origin in agents' own psyches.

Two types of cases come to mind, which we might call *reflexes* and *compulsions*. By 'reflexes', I mean actions that are not caused by desires, but those, such as the knee-jerk reflex, that are a pure physiological reaction to a stimulus.[41] By 'compulsions', I mean actions that are the result of irresistible desires, such as those experienced by sufferers of compulsive disorders or addictions. Such compulsions may be further subdivided by whether they express ill will or insufficient good will.

It is my contention that agents are not generally blameworthy for bad actions arising from reflexes. If a doctor stimulates my knee-jerk reflex and as a result my leg makes contact with him, such an action does not express ill will or a blameworthy lack of good will on my part. This action, if it can even be called an action, does not express any quality of will because it does not issue from any desire I have. It is no more indicative of my blameworthiness than the rate of fingernail growth or any other nonmental bodily process.

Compulsions are different in this regard. Compulsions are a type of desire and thereby do indicate one's quality of will. But we must be careful here, because acts of ill will and acts of insufficient good will are bad for different reasons. Agents are blameworthy for acts of ill will because these acts are *intended* to be harmful, whereas this is not the case for acts of insufficient good will. Rather, agents are blameworthy for acts of insufficient good will because they *could have done better*. Given that compulsive desires are difficult or impossible to resist, it may be the case that agents acting on such a desire could *not* have done better.[42]

Thus, the content of compulsive desires matters a great deal. If one has a compulsive desire to *cause harm*, then this desire expresses ill will – an intention to do wrong – and one would be blameworthy for acting on this desire, regardless of whether one could have done otherwise. We can make the analogous claim about praiseworthiness for compulsive acts of good will. If an agent acts from a genuinely good desire, then we justly hold such agents praiseworthy even if they could not have done otherwise. These 'volitional necessities'[43] are often experienced by emergency services workers, when they save lives and claim that they could not have done otherwise. These workers may not see themselves as praiseworthy ("I was just doing my job") but they undoubtedly are.

Compulsive acts of insufficient good will are different. Unlike compulsive acts of ill will, these are not intended to harm, and it therefore matters whether the agent could have done otherwise. In the case in which the compulsion is so strong that the agent could not have done otherwise, they cannot reasonably be blamed for their wrongdoing. Their action

[41] Dretske (1988)
[42] Arpaly (2005)
[43] See Watson (2002) for an extended discussion of volitional necessities.

does not express a *culpable failure* to do the right thing, any more than if they altogether lacked the ability to act on desires about others' mental states. Rather, such compulsive desires express an *inability* to act out of good will, although this inability is more localised than the general inability to act on desires about others' mental states.

That said, the term 'compulsion' is used for a range of desires, some of which are impossible to resist while others are merely unpleasant. Arpaly has claimed that an agent's blameworthiness for compulsive acts of insufficient good will differs by degrees, depending on how difficult it is to resist such desires. For instance, a drug addict who fails to break her addiction because the withdrawal symptoms are almost unbearable is less blameworthy than one whose symptoms are more mild.[44]

In practice, if an agent appears to act compulsively, it will often be difficult to determine whether they are blameworthy, because the content or the motivational strength of their desire will not always be obvious. While this poses the problem of determining an agent's blameworthiness, this is a practical problem rather than a structural one. The structural conditions for blameworthiness on my account are simple: agents are blameworthy for wrong acts of insufficient good will, provided they could have acted out of good will, and for wrong acts of ill will, regardless of whether they could have acted out of good will. If an agent meets one of these criteria for blameworthiness then they are blameworthy, regardless of any practical issues in determining that this is the case.

Conclusion to Chapter Two

In the previous chapter I argued that an agent's praiseworthiness or blameworthiness is best explained in terms of their ability to respond to the relevant reasons, and that this ability is constituted by an agent's ability to have desires about others' mental states. I claimed that this is a necessary condition of moral agency.

In this chapter I have argued that this ability is not just necessary for moral agency but also sufficient for it. In this discussion I have focused on the ability to respond to moral reasons itself rather than my analysis of this ability as constituted by the ability to have desires about others' mental states, but given my analysis in the previous chapter, I take the two abilities to be equivalent.

In arguing that the ability to respond to moral reasons is sufficient to be eligible for moral agency, I have considered additional conditions proposed by other philosophers as necessary for praiseworthiness and blameworthiness, and I have argued that these conditions are in fact not necessary. In many cases, I believe that these additional conditions are taken to be necessary because the philosophers in question have taken themselves to be giving an explanation not only of praiseworthiness and blameworthiness, but also of the justifiability of punishment, which I take to be a separate matter.

I've claimed that the ability to respond to moral reasons is the sole criterion for praiseworthiness and blameworthiness and is thereby sufficient for it. While I do not take myself to have given an exhaustive argument against all other possible criteria that one may think are necessary for praiseworthiness or blameworthiness, I do believe I've given good reasons to think the main ones – those that feature most commonly in the literature – are not necessary. I suspect that any other contenders will likewise be unnecessary for praiseworthiness and blameworthiness, primarily because Arpaly's account seems to capture all my intuitions about why praise and blame are warranted, but also because of

[44] Arpaly (2005)

the widespread (but by no means universal) view among philosophers working on moral responsibility that explaining moral responsibility necessarily involves explaining punishment.

So, there we have it. Given that moral agents are those individuals with the ability to meet the relevant responsibility conditions, and given that none of the conditions discussed in this chapter are necessary for agents to be appropriate targets of moral evaluation such as praise and blame, it seems plausible to conclude that basic moral agency requires only the ability to respond to moral reasons, which is in turn constituted by the ability to have desires about others' mental states. Compared with many other accounts of moral agency and responsibility, with their range of additional conditions, this sets a very low bar.

As such, we might wonder whether there is anything particularly special about adult human beings, any sense in which we are "better" moral agents. In the following chapter I will take up this question and argue that our ability to *guide* our behaviour by moral reasons marks a qualitative leap between moral agents of the more basic kind described so far and adult human beings as moral agents *par excellence*.

CHAPTER THREE: FLEXIBLE MORAL AGENCY

In Chapter One, I noted that moral agents have the ability to act morally but that there is an ambiguity in the phrase "act morally". This ambiguity concerned the level of cognitive access one has to one's behaviour. I distinguished between agents who (1) merely act in accordance with morality, (2) are motivated by moral reasons, and (3) use moral reasons to guide their behaviour. I claimed that merely acting in accordance with morality was insufficient for moral agency because one could do so purely accidentally, and that this would not warrant moral evaluation, such as moral praise.

The first two chapters were focussed on the second sense: being motivated by moral reasons. In Chapter One I gave an analysis of this ability in terms of an agent's ability to have desires about others' mental states and I argued that this ability was *necessary* for agents to be appropriate targets of moral evaluation, such as praise or blame.

In Chapter Two, I made the further argument that this ability is *sufficient* for moral agency. That is, I argued that agents' praiseworthiness and blameworthiness depend on nothing other than how they exercise their ability to act on the basis of (that is, to be motivated by) moral reasons.

In this chapter, I turn my attention to the third sense of "acting morally", using moral reasons to guide one's behaviour. As I mentioned in Chapter One, this is an important development over and above the simpler ability to be motivated by moral considerations. It is a *development*, because it builds on this simpler ability, and it is an *important* development because it allows for more sophisticated moral behaviour, including reason giving, self-improvement, and the teaching of morality to others.

I take this development to be so important that I think it is worthwhile to distinguish between two kinds of moral agency. *Basic* moral agency refers to the agency exhibited by individuals who are able to be motivated by moral considerations. On my analysis, basic moral agents are those who are capable of having desires about others' mental states.

Flexible moral agency is characterised by the ability to use moral reasons in order to guide one's behaviour. This ability requires significant explanation, which is the primary focus of this chapter. The secondary focus of this chapter is to explain how this ability allows for the sophisticated moral behaviour mentioned above.

To be clear, when I refer to basic moral agents, I am specifically referring to individuals who have the ability to be motivated by moral reasons but who lack the ability to guide their behaviour by these reasons. Flexible moral agents have the ability characteristic of basic moral agents in addition to the more sophisticated ability to use moral reasons in order to guide their behaviour.

Thus, basic moral agency represents for flexible moral agents a developmental stage between the non-agency exhibited by infants, and flexible agency, characteristic of agents such as ourselves. This development will be discussed in Chapter Four, in which I consider borderline cases of moral agency, such as young children.

The rest of this chapter is divided into four sections. In the first, I describe the distinction between motivating and normative reasons, and explore the relationship between these two types of reasons in the context of basic moral agency. I claim that praiseworthiness involve a correspondence between motivating and normative reasons, and that blameworthiness involves specific failures in this correspondence.

In the second section, I extend this analysis to flexible moral agency. In particular, I argue that guidance by moral reasons differs from mere moral motivation because the former involves beliefs about moral reasons, whereas the latter does not.

In the third section, I turn my attention to the types of moral behaviour afforded by flexible moral agency over and above those afforded by basic moral agency. I focus specifically on justification and improvement of agents' moral character.

Finally, I briefly consider how we ought to respond to flexible moral agents. I claim that their abilities, as outlined in the previous section, require us to hold them to a higher standard than that of basic moral agents, and that we should expect flexible moral agents to use these sophisticated moral abilities.

Motivating and Normative Reasons

When a person acts we explain their behaviour by reference to reasons. Philosophers distinguish between two types of reason, motivating reasons and normative reasons, which offer two different types of explanation.[1] Motivating reasons explain why a person acted by referring to the psychological states that were the cause of the action. That is, motivating reasons offer a *causal explanation*,[2] analogous to causal explanations in other domains, such as explaining salt dissolving in water by reference to the electrostatic forces of water molecules and salt ions acting on each other. The relevant psychological states that are typically taken to explain intentional behaviour are beliefs and desires. Just as electrostatic charges cause salt to dissolve in water, beliefs and desires cause people to act.

As a matter of semantics, a (motivating) *reason* for an action seems synonymous with the *explanation* of that action but also with the *cause* of that action. We could say that beliefs and desires are the reason for a particular action, or that they explain that action, or that they cause that action, and it seems that we would be saying the same thing in all three cases.

Typically, if I am asked my reason for performing a particular act, I will say that the reason *is* the cause of my act, not that the reason *refers to* this cause.[3] For instance, I might say that my reason for eating meat is that I want (that is, desire) to eat something that tastes good. I do not say that my reason refers to my wanting to eat something tasty.

But what then of explanations? Causal explanations simply *are* causes, not descriptions of causes.[4] I might say, for instance, that my desire to eat meat *explains* my doing so. The explanation, then, is my desire, which is a thing out in the world, not a description of that thing. I think this way of thinking is correct. Thus, the terms 'cause of behaviour', 'explanation for behaviour', and 'motivating reason' all refer to the same thing and I will use them interchangeably.

Turning now to normative reasons, these explain why a particular action was a good one to perform. Or, to be more accurate, normative reasons explain *at least in part* why an action *would be* a good one to perform. This specification is needed because normative reasons in favour of a certain action may be outweighed by normative reasons against the action, thus

[1] Smith (1994)
[2] Parfit (1997)
[3] Davidson (1963)
[4] Salmon (1989), Ruben (2003)

making the action a bad one to perform. For instance, the fact that meat is tasty is a normative reason in favour of eating it, but if this reason is outweighed by countervailing normative reasons, such as the fact that eating meat often involves animal suffering, then the action of eating meat would not be a good one to perform. That said, normative reasons, even for bad actions, are considerations that *count in favour* of that action;[5] they explain something about the action that counts in favour of performing it.

To give an example of a normative reason, say I were considering becoming vegan. The fact that this would prevent animal suffering is a normative reason for me to do so. Note that this normative reason exists regardless of whether I actually decide to become vegan. In general, normative reasons exist regardless of whether the person performs the action that they have a normative reason to perform.

That said, a major point of contention between philosophers working on normative reasons is whether they exist regardless of whether it's *possible* for the person to perform the action that they (purportedly) have a normative reason to perform. This possibility is generally understood in terms of whether the agent could be motivated to perform the action in question. A distinction is often made between *internalists* and *externalists* such that, internalists claim that a person has a normative reason to perform an action only if they could be motivated to do so, whereas externalists claim that a person may have a normative reason to perform an action even if they could not be motivated to do so.[6] For instance, if a psychopath could not possibly be motivated to care for other people, an internalist would claim that he has no (normative) reason to do so, whereas an externalist would likely disagree. Although I personally lean toward externalism, I have no stake in this debate for the purposes of developing an account of moral agency.

Normative reasons, unlike motivating reasons, are not causal explanations, because agents need not act according to their normative reasons. I have not become vegan even though I have a normative reason to do so; my behaviour is not explained by this normative reason. Instead, a normative reason is a *justification*: an explanation of what makes an action good to perform.[7] My normative reason to become vegan is the fact that doing so will reduce animal suffering.

Because normative reasons are not causal explanations, they are not desires.[8] But there is a relationship between an agent's desires and their normative reasons. Specifically, if an agent acts for a normative reason then the *content* of their desire refers to the normative reason in question. Suppose, for instance, that one were to become vegan. Now, there are many *motivating* reasons for such an action, some of which aim at good ends, while others do not. To take two very different reasons, one may want to prevent animal suffering, or one may want to annoy one's relatives at a barbeque. The prevention of animal suffering is a factor that counts in favour of becoming vegan, whereas annoying one's relatives does not. This is to say that the prevention of animal suffering is a normative reason to become vegan, whereas annoying one's relatives is not.

Thus, the agent who becomes vegan in order to prevent animal suffering is one whose motivating reason, the desire to prevent animal suffering, represents a normative reason, the prevention of animal suffering. Not all such agents are morally praiseworthy, since not all normative reasons are *moral* reasons. An agent who becomes vegan in order to be healthy is one who acts for a prudential reason rather than a moral one, but both prudential

[5] Scanlon (1998)
[6] McDowell (1995)
[7] Scanlon (1998)
[8] Smith (1994)

and moral reasons are a subset of normative reasons.[9] Moral reasons, as discussed in Chapter One, are primarily those that are concerned with the mental states of others, such as the animals whose suffering is avoided by one's choice to become vegan.

Praiseworthy agents, therefore, are characterised by a correspondence between their normative reasons – specifically their *moral* reasons – and their motivating reasons: their motivating reasons are desires whose content refers to their moral reasons.

However, this correspondence does not *wholly* constitute praiseworthiness, since it's possible for one to be motivated by moral reasons but nonetheless fail to act rightly. Consider the case in which I am motivated to become vegan by the desire to prevent animal suffering, and this desire causes me to buy food that (a) is vegan but which (b) indirectly caused more animal suffering than buying my preferred non-vegan alternative, such as buying food rich in unsustainably harvested palm oil instead of sustainably farmed honey. In this case, I was motivated to do the right thing for the right reasons but it's reasonable to suggest that I did not act rightly. Insofar as praiseworthiness requires rightdoing, and it seems to me that it does,[10] this behaviour is not praiseworthy.

Blameworthiness due to acts of ill will involves a similar correspondence, except in these cases, the correspondence is not between one's motivating and moral reasons. Instead, there is a correspondence between one's motivating reasons and what might be called one's moral *anti-reasons*. By this, and by normative anti-reasons generally, I am thinking of reasons that count *against* actions.[11] Having a moral anti-reason, say, to cause animal suffering, is the same as having a *moral* reason *not* to cause animal suffering. For agents who act out of ill will, this correspondence is structurally similar to that of agents who act out of good will. It involves a desire (the motivating reason) with content that represents the bad-making features of the act (the moral anti-reason).

Blameworthiness due to acts of insufficient good will does not involve such a correspondence between one's motivating reasons and one's moral anti-reasons, given that acts of insufficient good will are not motivated by the wrong-making features of an act. Acts of insufficient good will may be motivated by any number of considerations unrelated to the wrong-making features of the act. For instance, suppose I chose to eat meat simply because I like the taste. In this case, my motivating reason is my desire to eat tasty food. The content of this desire – the tastiness of the food – does not correspond to a moral anti-reason, since there's nothing intrinsically wrong with eating tasty food. It may correspond to several different alternatives. Firstly, it may correspond to a non-moral normative reason, such as a prudential reason. Secondly, it may correspond to a moral normative reason, but only a *pro tanto* reason,[12] since the purported wrongness of eating meat implies that this reason is outweighed by countervailing considerations. And thirdly, the content of this desire may correspond to no normative reason at all, if one holds that the tastiness of meat does not count in favour of our eating it. For the purposes of determining whether an agent is blameworthy for acting out of insufficient good will, it doesn't matter which of these alternatives corresponds to the content of their desire.

---

[9] Hare (1981)
[10] Arpaly (2002)
[11] Arpaly (2002) uses different terminology here. Instead of referring to *moral anti-reasons*, she uses the terms *anti-moral reasons* and *sinister reasons*. However, as indicated previously, I take reasons to be considerations that *count in favour* of an action, rather than those that count against. That said, the term *sinister reason* is one that is easily understood as referring to considerations that count against an action.
[12] Alvarez (2007)

What matters for blameworthy acts of insufficient good will, as discussed in Chapter One, is whether the agent *could have done better*. To express this idea in terms of motivating and normative reasons, it matters that the blameworthy agent had a moral reason to perform an alternative action *in the internalist sense*. This is to say that they *could have been motivated* to perform this alternative action for the relevant moral reason. For instance, suppose that eating meat is wrong because it causes animal suffering. But not all who eat meat are thereby blameworthy. For instance, I am blameworthy for eating meat but a dog is not.

Of course, if *externalism* is correct, then it's possible that both I and the dog have the same moral reason to refrain from eating meat: preventing animal suffering. But since the dog cannot be motivated by this reason, it cannot be blameworthy for failing to act on it. Note that I am not endorsing internalism here, since I am open to the possibility that the dog *has* a reason to refrain from eating meat. Rather, I am using the internalist conception of a normative reason (as one that could motivate agents to act) to identify an alternative action not taken by the agent, and for which the agent is blameworthy for failing to perform. Since I could be motivated to become vegan, I am blameworthy for failing to do so. Since the dog cannot be so motivated, it is not blameworthy for this failure.

There is of course an asymmetry here between agents who are blameworthy for acts of ill will and those who are blameworthy for acts of insufficient good will. I have claimed that the latter, but not the former, must have normative reasons in the internalist sense to act otherwise. That is, agents who act out of insufficient good will can only be blameworthy if they could have been motivated to do otherwise, whereas agents who act out of ill will can be blameworthy regardless of whether they could have been motivated to do otherwise. The difference, as discussed in Chapter One, is because the grounds for blameworthiness are different in the respective cases. Agents who act out of ill will don't need to have a reason to do better to count as blameworthy because they *intentionally* do the wrong thing. The intentional nature of their wrongdoing is what grounds their blameworthiness in these cases. Since agents who act out of insufficient good will do not intentionally act wrongly, they are only blameworthy if they *could have done better*. Such agents could only have done better if they could have been motivated to do so, thus requiring a normative reason (in the internalist sense) to count as blameworthy.

Guidance by Moral Reasons

If motivation by moral reasons is characterised by an accordance between one's motivating and moral reasons, what then of *guidance* by moral reasons? In other words, what does it mean to be guided by moral reasons and how is this different from mere motivation by moral reasons?

Let's begin with a simpler question. What is the difference between guidance and mere *causation*? Once we have this figured out, we can then return to the question of what it means to be guided by a moral reason.

If we were to say that an agent was motivated by something or other, we would be giving the *motivational reason* for their behaviour. This, as mentioned earlier, is a *causal explanation* of their behaviour. Of course, not all causal explanations involve motivational reasons. If I were to explain the behaviour of something other than an agent, such as the movement of a hot-air balloon, I would offer a causal explanation involving things such as heat and gravity, rather than one involving desires.[13]

[13] Dennett (1987)

Just as behaviour may be caused by things other than desires or reasons, behaviour may also be *guided* by other things. Consider the behaviour of a heat-guided missile. Unlike the hot-air balloon, whose upward lift is merely *caused* by the heat of its burners, the heat-guided missile responds in a more flexible way, sensing heat and moving toward it. How does it do this?

Since the missile can sense heat, then it must have a heat sensor, which gathers information about the heat in the surrounding environment. And since the missile can move toward the heat, it must use this information to determine where to move. That is to say, the missile *represents* the heat in the environment and uses this representation to guide its behaviour. Compare this to the behaviour of the hot-air balloon, which also uses heat to move, but does so directly, by heating the air in the balloon, rather than indirectly via the use of a representation of heat. It is this use of representations that distinguishes guidance from mere causation.

Consider another pair of examples: a ball moving through the air and a sportsperson throwing a ball through the air. The movement of the ball is directly affected by the physical forces acting on it, including the force of the throw and the pull of gravity. The movement of the sportsperson, by contrast, is largely caused by his use of representations: his belief that if he throws the ball just so then it will travel in the desired way. His belief is informed by his understanding of how the force of his throw and the force of gravity affect thrown objects. Thus, the movement of the ball is caused directly by physical forces, whereas the movement of the sportsperson is caused at least in part by mental representations of these forces.

In general, the difference between guidance and mere causation is that the former involves a representation, whereas the latter typically does not. Guidance by heat, or gravity, or the force of a throw, is characterised by the use of representations of these factors, whereas mere causation by these things does not involve a representation. However, the analysis is complicated when considering causation *by reasons*. This is because causation by reasons, which is to say motivation, involves desires, which are representations.

The distinction is further complicated by the fact that one can be guided by different kinds of reasons. In particular, guidance by motivational reasons differs from guidance by normative reasons, because guidance in general involves representation of the thing doing the guiding, and motivational and normative reasons are different kinds of things. Motivating reasons are desires, so guidance by motivating reasons involves representation of desires. An example of guidance by a motivating reason is thinking ahead in a game of chess. If I know that you intend to take my queen in two moves then I represent your desire to do so and use this to guide my game strategy. Guidance by motivating reasons is important when dealing with other agents, but it is not directly relevant to the idea of guidance by moral reasons, since moral reasons are a type of normative reason.

By contrast, normative reasons are not desires, but considerations that count in favour of something. Guidance by normative reasons, then, involves representation of these considerations. For example, if I am thinking about becoming vegan because I dislike animal suffering, then my thoughts about this are mental representations whose content refers to animal suffering. Unlike guidance by motivational reasons, they do not seem to refer to any *desires* about animal suffering.[14]

Given that moral reasons are normative reasons, we have a problem. *Motivation by normative reasons* involves representation of the considerations that count in favour of the action,

[14] Smith (1994)

because such motivation involves acting on a desire which represents the normative reason in question. If I am motivated to become vegan because I desire the prevention of animal suffering, then this desire represents the relevant normative reason. But, as we have just seen, *guidance by normative reasons* also involves representation of the considerations that count in favour of the action, so the mere fact that guidance involves representation does not here distinguish motivation and guidance by normative reasons, including moral reasons.

One suggestion for distinguishing guidance from mere motivation is to claim that the former requires conscious deliberations,[15] whereas the latter does not. There is some plausibility to this view. If I am consciously deliberating about becoming a vegan, then I will consider the reasons in favour and use these reasons to guide my decision. But while conscious deliberation may provide evidence of guidance, it seems that one can be guided by something without consciously deliberating about it. The heat-guided missile, for instance, is guided by heat but does not consciously deliberate about it. I cannot think of a reason why this wouldn't apply to guidance in general, including guidance by moral reasons. For instance, suppose I became vegan not because I deliberated about the reasons for doing so but because I watched a documentary about factory farming and felt revulsion at the animal suffering depicted in the documentary. It seems that in this case, I am guided by the relevant moral reason to become vegan without consciously deliberating about it.[16]

Guidance by moral reasons, therefore, needs something more than mere *representation* of the moral reasons, since this cannot distinguish it from mere motivation by moral reasons, but this 'something more' is not conscious deliberation about the relevant reasons.

The correct answer, I think, is that guidance by moral reasons involves *moral judgements*.[17] When I watch the documentary, I form a moral judgement that eating meat is wrong, even though I did not come to this judgement via conscious deliberation. It strikes me that moral judgements are *beliefs* about the moral properties of certain things.[18] For instance, if I judge animal suffering to be wrong, I have a belief about animal suffering. In particular, I have a belief of the form 'suffering is wrong', wherein I ascribe the property of wrongness to this suffering.[19]

This is to say that moral judgements differ structurally from moral motivation in two ways. Firstly, judgements are beliefs, rather than desires. Secondly, and more subtly, the representational content of moral judgements differs in an important way from that of moral motivation. As discussed in Chapter Two, illustrated by the case of Huckleberry Finn, an agent may be motivated to do the right thing without realising that his actions are right;[20] moral motivation requires representation only of the good-making features of the act. Moral judgement differs in that agents must represent both the good-making features *and* the fact that these good-making features are in fact good.[21] This is what I mean when

[15] Hacker (2007), Korsgaard (1997)
[16] Arpaly (2002, 2006)
[17] Smith (1994)
[18] *Ibid.*
[19] This thereby commits me to moral cognitivism, but I think this is the most plausible commitment to make with respect to the properties of moral statements. A defence of this commitment would take me beyond the scope of the thesis, but I am persuaded here by the arguments of Michael Smith (1994).
[20] This is often referred to as the distinction between rightdoing *de re* and *de dicto* (Arpaly (2002)), such that one acts rightly *de re* if they do what is in fact right, and they act rightly *de dicto* if they do what they believe is right.
[21] van Roojen (2018)

I say above that a judgement is a belief of the form 'X is wrong', wherein one ascribes the property of wrongness to X.

But how does one represent the fact that good-making features are good? How does one represent the concepts of goodness and badness, of rightness and wrongness? The answer, I think, is by representing the *evaluative properties* of these good-making features.

Of course, not all evaluative properties are *moral* properties, for the same reason that not all value judgements are moral judgements. For instance, one may make *taste* judgements about, say, the flavour of vegan food. Unlike the moral judgement that animal suffering is bad, this is not a moral judgement for or against becoming vegan because it is not other-regarding, whereas moral judgements are. This suggests that the kinds of evaluative properties that count as moral properties are those that are about others. As I argued in Chapter One, these evaluative properties are specifically about others' mental states. Thus, the badness of animal suffering is a moral reason, given that the reason is specifically about the mental states of other beings.

Putting this together, I have claimed that moral judgements are beliefs about evaluative properties of others' mental states. The ability to form such beliefs requires the ability to form beliefs about (a) evaluative properties and (b) others' mental states. Of these, the ability to form beliefs about others' mental states is the more sophisticated ability, often known as *theory of mind*,[22] a topic to which I will return in Chapter Four. The ability to form beliefs about evaluative properties is simpler. If I believe that meat is *better than* vegan alternatives, then I have a belief about an evaluative property.

We might call such beliefs *comparative beliefs*, because they involve comparing the value of some things to others. It is important to note that comparative beliefs are not the same as mere *preferences*, since preferences need not be represented by beliefs.[23] A preference for something *may* involve a belief that it is better than an alternative, but it may instead involve a desire for this thing over an alternative. That said, the ability to *express* this preference does seem to involve a comparative belief.

We can have comparative beliefs about many things, but it strikes me that guidance by moral reasons involves comparative beliefs about desires. For instance, if I judge that I ought to become vegan because doing so would reduce animal suffering, then I am judging that I ought to act on the desire to prevent animal suffering rather than some other desire. That said, not all comparative beliefs about desires constitute moral judgements, because not all desires are about the mental states of others. If I believe that I ought to act on my desire to eat meat instead of my weaker desire to eat vegan alternatives, I am making a value judgement but not a moral judgement. Thus, guidance by moral reasons involves a particular kind of moral judgement, a comparative belief about desires, which are in turn about the mental states of others.

An objection could be raised here. If moral judgements necessarily involve comparative beliefs about desires, this seems to conflate moral goodness with *desirability*, whereas some things are good without being desirable. One might think that Nelson Mandela's goodness, for instance, was due to his virtuous character, which is admirable but not necessarily desirable. Good things may be desirable, but they may also be admirable, or enjoyable, or so on. [24] I think this is a perfectly fine response to the claim about what constitutes

[22] Doherty (2007)
[23] Heathwood (2014)
[24] Nozick (1974)

goodness *in general*. Indeed, the goodness of moral agents makes them praiseworthy, which is similarly distinct from desirability.

That said, the fact that praise is the specific fitting attitude one ought to have toward goodness in moral agents suggests that there might also be a specific fitting attitude toward the kind of goodness relevant to normative reasons for action. Since this type of goodness is supposed to guide our behaviour, it makes sense that this is the type of goodness that we ought to be *motivated* to bring about. That is, this kind of goodness is such that we should *desire* it; it is *desirable*.

This is a clear structural difference between moral motivation, characteristic of basic moral agency, as developed in Chapter One, and moral guidance, characteristic of flexible moral agency in the sense developed here. Moral motivation requires desires about others' mental states, whereas moral guidance requires representation of a higher order: comparative beliefs about desires about others' mental states. This ability to *metarepresent* will be discussed in detail in the following chapter. For now, though, we turn our attention to the actions enabled by moral guidance over and above those enabled by mere moral motivation.

Justification and Moral Improvement

The ability to guide one's behaviour by moral reasons represents a significant development beyond the simpler ability to merely be motivated by moral reasons. In particular, it involves *using moral reasons*. By 'using moral reasons' I mean exactly the ability outlined in this chapter: such agents mentally represent moral reasons, and these representations play a role in their behaviour (in contrast to moral reasons simply motivating behaviour in the absence of an intermediary representation of the reason).

In this section, I will make a necessity claim but not a sufficiency claim. That is, I claim that the ability to use moral reasons is necessary, but may not be sufficient, for several commonplace forms of moral activity. Central to these activities is *justification*, by which I mean the act of communicating reasons, usually to others but also to oneself. Acts of communication require a mental representation of the thing being communicated. For instance, if I wish to tell someone that there is milk in the fridge, then I must mentally represent the milk in the fridge in order to communicate this. Similarly, if I wish to tell someone that I used the last of the milk in my daughter's bedtime bottle, then I must mentally represent this reason in order to communicate it. Without the ability to mentally represent reasons, one cannot use these reasons to justify one's actions, or the actions of others.

Justification, Excuse, and Apology

Justification is typically held to be distinct from excuse, insofar as the former involves denial of wrongdoing while the latter involves denial of blameworthiness.[25] A justification for an act of apparent wrongdoing would involve giving a reason why the act was not in fact wrong. An excuse would accept that the act was wrong, but would involve giving a reason why one should not be blamed for the act of wrongdoing. Note in both cases, however, that one gives a reason for one's behaviour. If these are both *moral reasons*, then

[25] Wallace (1994)

this would imply that the ability to use moral reasons is fundamental to both justification and excuse.

Consider a paradigm case of justification: explaining a case of triage, such as that of a field medic allocating scarce medical resources so as to maximise the chance of the most patients surviving, but which resulted in the deaths of some patients who would have survived if the resources were allocated differently. To explain this action, one would point out that the chosen action was the least-worst option from several undesirable alternatives. The fact that 'least-worst' here refers the relative numbers of survivors, and that more survivors is better because it generally involves less suffering, may be assumed to be common knowledge, but if this were in doubt then these facts would also be explicitly mentioned as part of the justification. Justification of this sort – moral justification – clearly involves the communication of moral reasons.

Unlike justification, when one makes an excuse, one does not give reasons for the rightness of one's actions. Because of this, one may be inclined to think that that excuse does not involve giving moral reasons. I think this is mistaken. Consider a case in which a child accidentally but carelessly injures his classmate. The child may seek to excuse his behaviour by claiming that it was just an accident. Unlike in cases of justification, the child does not claim that injuring his classmate was the right thing to do. Rather, the excuse serves to deflect blame for his action. It may even be the case that the child does not even care whether it was right or wrong to injure his classmate, only that he not be blamed for it.

In such a case, how can giving an excuse be an instance of giving a moral reason? Consider that in claiming that the injury was accidental, the child makes an implicit distinction between types of behaviour that exempt one from blame and those that do not. Generally, making an excuse in order to avoid blame reflects an understanding of this distinction. The most plausible way of making this distinction, as I have argued, is that blameworthy wrongdoing involves acting wrongly for the very reason that makes that act wrong (acting out of ill will) or acting wrongly despite being able to act rightly for the right reasons (acting out of a culpable lack of good will). This is distinguished from blameless wrongdoing wherein one acts wrongly because one was unable to have acted rightly for the right reasons, either by being unable to act rightly at all or by being unable to perceive the relevant reasons for the alternative right action. If one understands the distinction between blameworthy and blameless wrongdoing in this way, then offering an excuse to deny blameworthiness is a clear use of moral reasons.

However, one need not understand blameworthiness in this way in order to deny blameworthiness for one's wrongdoing.[26] That said, in doing so, one must have *some* criterion by which to distinguish blameworthy from blameless wrongdoing. And unless one conceives of blame as wholly separate from wrongdoing, such that wrongdoers are generally no more blameworthy than right-doers, then it strikes me that the distinguishing criterion must relate in some way to the relevant moral reasons that distinguish wrongdoing from right-doing. For instance, when a child claims that an injury was accidental, it seems that part of the intended explanation was that the injury wasn't *intended* and that such an intention, if it were present, would be grounds for blameworthiness. This in turn seems to demonstrate an understanding that *it would be wrong* to act on such an intention, that there would be a *reason* not to do so. In offering "it was just an accident" as

[26] Sher (2009)

an excuse, it seems that the child is using some reason about the action's wrongness as a reason to avoid blame, but is this a *moral* reason?

If the reason is that injuries hurt people, then this is clearly a moral reason. But it's conceivable that the child is unaware that injuries hurt people or that this is a reason not to injure others, and that he just thinks it is wrong because his parents or some other authority told him it was wrong. If the child thinks that "because an authority figure said so" is the only reason not to perform some act, then this does not seem like a moral reason. And if this child used "it was just an accident" to avoid blame, then it strikes me that this child would not be using a moral reason when offering an excuse for their behaviour.

Moreover, excuses can also be a learned response to the unpleasantness of being blamed, particularly for very young children. In these cases, the response "it was an accident" need not indicate un understanding of the relevant reasons, but excuses of this kind can be easily distinguished from 'genuine' excuses on the grounds that they are not given as a *reason* for avoiding blame but as a *reaction* to an aversive stimulus. (That said, these 'reactive excuses' cannot have their desired effect unless they are given in a social context in which genuine excuses already operate. So, even though agents may be able to give reactive excuses without being able to use reasons, these excuses depend on other agents giving moral reasons in order to function.)

The upshot is that while offering excuses *typically* involves the use of moral reasons, it is not *necessary* that they do so, particularly when the person giving the excuse is a very young child, and thus excuses do not require flexible moral agency in the way that justification does.

Apology is different again from justification and excuse. Unlike justification, apology does not deny wrongdoing, and unlike excuse, it does not deny blameworthiness. Genuine apology admits both wrongdoing and blameworthiness and seems primarily an attempt to restore relationships damaged by such.[27] Even cases of non-genuine apology, apologies of the form "I'm sorry you feel wronged by my action" are attempts to restore relationships, although they seem also to sidestep admissions of blame and wrongdoing, rather than denying such admissions altogether. What is interesting about cases of both genuine and non-genuine apology is *how* they attempt to restore relationships. Unlike, say, grooming behaviour in chimpanzees, which can also serve to restore damaged relationships, there is a clear moral dimension to the practice of apology.

Apologies, even non-genuine ones, are apologies *for* perceived acts of wrongdoing. When I say 'perceived' acts of wrongdoing, I mean to include acts that are perceived as wrong by the person to whom the apology is directed, even if the person apologising disagrees, as well as acts that are perceived as wrong by the person apologising, even if the person to whom the apology is directed disagrees. Both types of case involve a perception of an act *as* an act of wrongdoing. By this I mean that any specific act might be perceived as wrong or not – for instance, an act of theft may or may not be perceived as wrong – but when one apologises, the act *is* perceived as wrong, either by the apologiser or the apologisee or by both. In perceiving an act as wrong, one mentally represents both the act and its purported wrongness.

As we have seen in the case of excuses, perception of an act as wrong typically involves an understanding of *why* the act was wrong. Typically, but not always. In apologising for causing an injury, for instance, one may think that it was wrong to injure another because

[27] Bennett (2008), Dunbar (1998)

of a moral reason, such as the fact that injuries harm people, or one may think that it was wrong because of a non-moral reason, such as the fact that injuring others is prohibited by the relevant authority. Given this, it does seem possible to apologise without using moral reasons, although these cases seem to rely on a failure to understand the reasons for which a wrong act is wrong.

Earlier I mentioned a distinction between genuine and non-genuine apology. It is my contention that this distinction is wholly independent from the distinction between apologies that involve the use of moral reasons and those that do not. That is, some cases of genuine apology involve the use of moral reasons and some do not, and some cases of non-genuine apology involve the use of moral reasons and some do not. To see this, first consider that genuine apology is characterised by 'meaning it', where 'it' seems to refer to the belief that one's actions were in fact wrong.[28] Thus, genuine apology requires not only a representation of one's act as wrong, but that one takes this representation to accurately reflect reality. However, this is independent of the use of moral reasons, since one may believe that one's actions were wrong for a non-moral reason, such as the actions being prohibited by an authority, and offer a genuine apology on these grounds.

Thus, while justification involves the use of moral reasons, not all cases of excuses and apology do so, which implies that moral agency is not necessary for excuses and apology. That said, justification is central to our moral lives. As I shall discuss below, it plays an important role in the improvement of our moral character.

Moral Improvement

The development of moral agency is not a switch. It's not as if children become moral agents one day and remain the same way for the rest of their lives. People usually become *better* moral agents over time – more compassionate, more considerate, and so on. Moreover, people often help others to become better moral agents. I shall refer to the ability to help ourselves and others become better moral agents as *moral improvement*. I contend that while moral improvement of behaviour is possible without the ability to use moral reasons, *reliable improvement of moral character* is not.

The distinction between behaviour and character is an important one, and requires a little elaboration. Improvements in moral behaviour are indicated by increasingly better evaluations of agents' *actions*. Over time, agents will perform fewer wrong actions and more right actions. By contrast, improvements in moral character are indicated by better evaluations of agents' *quality of will*. Over time, agents will less often be blameworthy and more often be praiseworthy. In general, this can only occur if agents act for increasingly better reasons, or if they act for good reasons more often. To do this *reliably*, as opposed to accidentally or haphazardly, requires the ability to distinguish these reasons from bad ones, which is to say it requires the ability to have beliefs about moral reasons.

To see this in context, let's consider three methods of moral improvement: direct communication, role modelling, and discipline. By *direct communication*, I am thinking primarily of cases wherein one person tells another that they ought to act or refrain from acting in a certain way, such as telling one's child that they ought to share their toys. By *role modelling*, I am thinking primarily of cases wherein a person's behaviour is intended to be imitated by another, such as refraining from yelling in front of one's children in order to demonstrate respectful behaviour. By *discipline*, I am referring to the practice of imposing adverse consequences for misbehaviour, such as sending a child to their

[28] Joyce (1999)

bedroom or banning them from using electronic devices. I also have in mind corporal punishment,[29] which is often intended to improve moral character, regardless of whether it is in fact effective in this regard.

Each of these cases involve an actor – the person doing the communication, role modelling, or discipline – and a target – the person to whom the communication, role modelling, or discipline is directed. In this discussion, I will use parents and children as examples of actors and targets, respectively. That said, the actors and targets need not be parents and children; they may even be the same person, given that we can use these methods to improve our own moral character. However, in cases of other-directed moral improvement, such as those involving parents and children, there is the additional question of *who* uses moral reasons. It is my contention that improvement of moral character can reliably occur provided that *either the actor or the target* uses moral reasons.

Let's consider direct communication. Consider the case in which a parent asks their child to share their toys with their sibling. This is a fairly straightforward case of direct communication: a directive aimed at eliciting moral behaviour. As currently described, it does not necessarily involve the use of moral reasons, since no justification is given and perhaps the parent has no justification in mind. Nonetheless, directives are often accompanied by moral justifications. For instance, the parent might add that sharing will make our siblings feel better. In doing so, the parent issues a directive and offers a moral justification, which necessarily involves the use of moral reasons.

Alternatively, the parent might say that if we don't share with others then others won't share with us. This is less obviously a moral justification, as this reason could be understood as prudential.[30] Even so, by encouraging children to act in their rational self-interest, parents communicate their concern for their child's interests, which again involves the use of moral reasons.

Or the parent could say that the child should just obey, or omit any justification whatsoever. Assuming, as seems reasonable, that mere obedience is not a moral reason, [31]

[29] The mention of corporal punishment raises the question of the difference between punishment and discipline. Although both involve the imposition of adverse consequences in response to misbehaviour, there seems to be a difference in their respective *aims*. Punishment strikes me as fundamentally retributive. Targets of punishment are taken to *deserve* their punishment: the pain is the point, and it is irrelevant whether the target becomes a better person as a result. Discipline strikes me as fundamentally aimed at moral improvement. If it is effective, then the target of the discipline will become a better person, or will at least become better behaved: the pain is a means to this end. Of course, the two often occur together. In many cases, people aim to discipline *and* punish at the same time. Nonetheless, they are different practices with distinctly different aims.

[30] Nagel (1986)

[31] I need to be careful here. I do not wish to claim that obedience to authority is never an *instrumental* moral reason. A warehouse worker for a charitable organisation may do a lot of good by following directions to ensure that boxes of food are labelled with the correct shipping destinations, for instance. What I wish to say is that obedience to authority is not an *intrinsic* moral reason to perform any action. This runs contrary to some views, for instance Haidt's (2012, p.144-148) claim that obedience to authority is one of six moral foundations, alongside more traditional foundations such as the three I discuss in chapter one (concern for welfare, autonomy, and fairness). My objection to obedience to authority as an intrinsic moral reason rests on a Euthyphro dilemma: either obedient actions are right because the relevant authority has an independent moral reason for their directions, in which case obedience is an instrumental reason, or obedient actions are right purely because the relevant authority says so, in which case the directions issuing from the authority are arbitrary. In the example above of the parent asking for obedience, I assume that they don't have an independent moral reason for their direction (otherwise this would involve use of a moral reason for judging the disobedient behaviour as wrong), and thus that their direction is arbitrary.

a few things might happen here. Firstly, it's likely that the parent is aware of some other moral reason for asking their child to share. In this case, the parent still uses a moral reason but this reason remains unexpressed. Secondly, it's least possible that the child might supply their own moral reason. It may dawn on the child that they're being asked to be obedient whenever their disobedience causes hurt feelings. Here too, moral reasons are being used, but by the child rather than the parent. In both cases, we might expect the child to become better at sharing due to the use of these moral reasons.

But suppose, thirdly, that there is no use of moral reasons by either the parent or the child. Suppose that the parent has no implicit moral reason in mind when asking their child to obey and that the child comes to believe that she ought to obey her parents, or that obedience is a good thing more generally. It is possible in such cases that one may end up committing wrongful acts out of obedience. The phrase "I was following orders" may be uttered as a justification for one's actions, but it does not typically indicate that one's action was right.

That said, it is possible for one to improve one's moral behaviour without using moral reasons to facilitate this improvement. If a parent were to request obedience disproportionally in cases where such obedience would have a morally good outcome, such as asking children to share, and if the child were never to realise that there were moral reasons for obeying these directives, the child could become better at sharing without anyone using moral reasons. But even if one's obedient actions are morally right, one would not be *praiseworthy* for these actions, as they were performed not for the relevant moral reason but merely out of obedience. A child who acts for this reason is no more praiseworthy than a grocer who prices his goods fairly in order to maximise his profits. Such obedience could allow for accidental moral improvement, say if obedience to authority or prudence were reliably correlated with improved moral outcomes, but even if such correlation were the case, this would only result in improvements in one's *behaviour*, not one's moral *character*. Reliable improvements in moral character seem to rely on the use of moral reasons. Direct communication of these moral reasons is one way of using these moral reasons to reliably improve moral character.

But it is not the only way: an alternative is to imitate a virtuous exemplar. If such imitation is to result in moral improvement, one must choose the right role model, since choosing a role model arbitrarily may instead result in moral deterioration. The most obvious criterion for choosing the correct role model is right action; a role model who consistently does the right thing is one worth emulating.[32] For this, one needs to be able to recognise which actions are right, which in turn requires the ability to recognise the *right-making features* of actions. If one is unable to do this, then one may end up misidentifying which actions are right. As discussed above, one may mistakenly believe that obedience to the relevant authority is the right-making reason for most actions, and one may imitate a role model who exemplifies this trait, with the result that one may commit wrongful acts out of obedience. Thus, in order to reliably choose a good role model, one must have some understanding of moral reasons.

That said, children do not *choose* all of their role models, particularly their parents, who are typically their first role models. Moreover, this role modelling begins before children are capable of using moral reasons. And yet, children generally become better moral agents as they mature. An obvious reason for this is that *their parents* often use moral reasons throughout the process of parenting. We have already considered parents' use of direct

[32] Aristotle (350BCE/1951)

communication of moral reasons. Often, parents will also use moral reasons in other ways when role modelling.

Sometimes they do this knowingly. If a parent wants their children to be respectful to others, they may make an extra effort to model respectful behaviour in front of their children. They may reflect on their own behaviour and resolve not to yell at their children. In doing so, parents recognise that it is more respectful to speak calmly than to yell, and in recognising this, they use moral reasons.

But not all role modelling is explicit. For instance, we are typically quiet when we visit places such as libraries, cinemas, and restaurants. This is also true much of the time when parents visit these places with their children. While this sometimes involves explicit role modelling of the desired behaviour, it is often habitual. Nonetheless, children notice the difference in volume and often become quieter themselves. In this case, habitual role modelling results in moral improvement in the child's behaviour, as being quiet in such places makes a more pleasant experience for other visitors to these places.

It is at least conceivable that this kind of implicit role modelling need not involve the use of moral reasons, since the parents could have picked up the habit of being quiet in these places from their own parents, who could have picked it up from their own parents, and so on, without anyone reflecting on why they ought to be quiet. However, and as is also the case for direct communication, any moral improvement that occurs without the use of moral reasons will be limited to improvement of behaviour, rather than of moral character.

It is more plausible, though, that parents develop the habit of being quiet in quiet places by various means, including both implicit and explicit role modelling, and direct communication of both desired behaviour and the reasons for this. It is plausible that at least some of these means involved the use of moral reasons, so even if these parents do not *themselves* use moral reasons when role modelling in front of their children, their doing so is likely the result of moral reasons being used at some point.

A third method of improving both moral behaviour and moral character is discipline. As with direct communication and role modelling, discipline may be administered without the use of moral reasons, for instance as an emotional reaction to misbehaviour. But, again, unless discipline involves the use of moral reasons, it will not result in reliable improvement of moral character.

Consider that if discipline is to be effective in improving moral character, then it needs to play a role in motivating the target of the discipline to act for the right reasons in the future. For this to happen reliably, there needs to be a causal connection between the discipline and the disciplined person's reasons for their future actions.[33] How might such a causal connection be made? Ideally, the discipline would give the disciplined person the opportunity and incentive to reflect on why their action was wrong so as to avoid acting for this reason in the future. On this picture, if a child is disciplined for an act of wrongdoing, the child will then come to understand that their behaviour was wrong and should be avoided in the future. This understanding may occur as a result of their consciously deliberating about the reasons their action was wrong, but it may also occur by some nondeliberative process, such as the relevant reason dawning on them later. In either case, the subsequent moral improvement is attributable to the disciplined person's use of moral reasons.

[33] Fischer (2004)

Of course, this is the ideal case. For any act of discipline, it is at least as plausible that it does not prompt the child to reflect on the relevant moral reasons, or to have these reasons dawn on him. In these cases, it is still possible for discipline to have the desired effect; for instance, discipline can act as an aversive stimulus, causing the child to act correctly in future by a process of behavioural conditioning. As we have seen with both direct communication and role modelling, this is most likely to occur if the parent uses moral reasons in the process of disciplining their child.

There are two places where this may occur. Firstly, by disciplining their child for a perceived act of wrongdoing, this indicates that the parent judges the behaviour to be wrong, thereby showing that they distinguish between right and wrong acts. Secondly, it also indicates that the parent believes that discipline is an appropriate response to this behaviour, thereby showing that they distinguish between appropriate and inappropriate responses, or again, right and wrong acts. At two separate points in the process, the parent relies on a distinction between acts that are perceived as right and those that are perceived as wrong, a distinction that requires a distinguishing criterion. As I previously claimed, this criterion need not be an actual moral reason since one may be mistaken about which reasons are moral reasons and which are not. But unless the child independently uses actual moral reasons to change their future behaviour, the discipline is unlikely to result in moral improvement, just a more obedient child.

I have discussed three methods of improving moral character: direct communication, role modelling, and discipline. In all three cases, I have claimed that reliable improvement cannot occur without the use of moral reasons. It is not necessary that the person engaging in the relevant practice uses moral reasons, provided that the person to whom it is targeted does so. If neither party uses moral reasons then at best, such improvement will be either accidental or limited to improvements in behaviour, rather than character.

In these examples, I have focused on the use of moral reasons as a *selection mechanism*. That is, the ability to use moral reasons allows agents to select which action they ought to perform. For instance, by telling a child that they ought to share because doing so will make their sibling feel better, the parent offers the child a reason to choose to share rather than choosing not to share.

But the use of moral reasons can improve moral character in a second way: as a *generation mechanism*. Consider that we may be praiseworthy for many actions without such actions constituting moral improvement. For instance, if a child is predisposed to act out of concern for others and freely shares her toys with her sibling, such behaviour is surely praiseworthy, but future acts of sharing are a continuation of this praiseworthy behaviour, rather than constituting moral improvement over her already praiseworthy self. For an agent's moral character to *improve*, there must be some change in their desires. For instance, if the child was initially unconcerned with her sibling's feelings, but later came to desire happiness for her sibling, and this desire motivated her to share with her sibling, then this act of sharing would constitute a moral improvement. In this case, a new desire is *generated*, which motivates her to perform an action for which she is praiseworthy.[34]

[34] The generation of a new desire is not the only way in which agents' moral character may improve. For instance, the child may already desire their sibling's happiness but not enough to overcome her desire to keep her toys for herself. In this case, a change in the relative strengths of these desires may cause the child to start sharing, for which she would be praiseworthy. Alternatively, she could lose this competing desire altogether, which may result in similarly praiseworthy behaviour. Moreover, each of these changes – generating new desires, losing existing desires, and changing

How might one come to generate such a desire? Realistic accounts have overwhelmingly focused on moral judgements as the cause of moral behaviour.[35] As discussed above, moral judgements involve the use of moral reasons. There are several different accounts of how moral judgements affect our desires, and the primary difference between them is the strength of the connection between moral judgements and moral motivation.

Very few accounts posit a *necessary* connection between moral judgement and moral motivation, given that we are so often beset by weakness of will; for instance, a child's judgement that she ought to share need not motivate her to share.[36] Nonetheless, I am not aware of any account in which there is no connection whatsoever between moral judgement and moral motivation.

The primary dispute seems to be between *Humeans*, who claim that moral judgements can only motivate agents by way of some pre-existing desire, and *anti-Humeans*, who claim that moral judgements can motivate agents independently of their pre-existing desires.[37]

My contention that moral judgements generate desires with moral content strikes me as a truism for anti-Humeans[38] but more difficult to reconcile for Humeans. Nonetheless, I think that even Humeans can readily accept this claim. To see this, consider how existing desires might give rise to new ones. If I desire a coffee and I believe that the café serves coffee, then I could form the new desire to go to the café. This new desire is derived directly from my existing desires.

Similarly, if a child has a pre-existing desire to be happy, and then forms the moral judgement that others' happiness is relevantly similar to her own, this could give rise to the desire to make others happy. Thus, Humeans can accommodate the claim that moral judgements give rise to desires with moral content, provided that they do so in conjunction with an agent's pre-existing desires.

This is to say that moral judgements can generate new desires, and we can accept this regardless of whether we accept a Humean or anti-Humean account of moral motivation. By generating these desires,[39] moral judgements can cause improvement in moral character. Although I have not shown that moral judgements are the *only* way of generating new desires, I am not aware of any plausible alternative.[40]

---

relative motivational strengths of desires – can result in moral improvement not just by making agents more praiseworthy, but also by making them less blameworthy. Nonetheless, while the subsequent discussion will focus on the role of moral judgement in desire *generation*, the relevant claims generalise to cases of desire *alteration* more broadly, including both the loss and change in motivational strength of desires.

[35] I specify *realistic* accounts here, as there is a large portion of the responsibility literature focused on manipulation arguments, which involve evil neurosurgeons 'implanting' desires into unwitting victims. Of course, such cases are not intended to be realistic accounts of how desires are generated.

[36] Mackie (1977) seems to be an exception here.

[37] Rosati (2016)

[38] That is, this strikes me as a truism provided that anti-Humeans accept that moral judgements motivate agents by *generating the relevant desire*, rather than motivating agents *directly*, without inducing any desire. Of course, some anti-Humeans claim that moral judgements can motivate directly (e.g. Shafer-Landau 1998). Given my commitment to belief-desire psychology, as outlined in Chapter One, I find this view implausible, but it would be beyond the scope of this thesis to critique such accounts in detail. For a fuller defence of the claim that moral judgements are beliefs, not desires, and do not motivate directly, see Smith (1994).

[39] As well as by other alterations to desires, as noted in Footnote 35.

[40] An alternative that is sometimes suggested (e.g. by Schroeder (2004) and Sinhababu (2017)) is that some desires emerge as a natural result of psychological development. The adolescent's desire for sex, or indeed the toddler's desire to do things for oneself, are examples. But this leaves open

As we have seen, the ability to form moral judgements allows flexible moral agents to improve their moral character and that of others. This ability arises from two mechanisms: a selection mechanism, wherein the moral judgement helps us to choose how to act, including our participation in practices such as direct communication, role modelling, and discipline, and a generation mechanism, wherein these judgements give rise to new desires and changes in our existing desires. Without the ability to form moral judgements, it seems very difficult, perhaps impossible to improve one's moral character. In the following section, we shall consider the implications of this.

Normative Expectations of Flexible Moral Agents

Now we return briefly to the broad definition of moral agency introduced in Chapter One: moral agents possess certain *abilities*, which enable certain *actions*, the performance of which opens agents up to certain kinds of *response*.

For basic moral agents, the relevant ability is the ability to respond to moral reasons by acting on desires about others' mental states; the relevant action is the non-accidental performance of morally right and morally wrong actions; and the relevant response is moral evaluation of these agents as praiseworthy or blameworthy.

For flexible moral agents, the relevant ability is the ability to be guided by moral reasons by using mental representations of these reasons as moral reasons, and the relevant actions are those discussed in this chapter: justifying actions and improving moral character. I have not yet said anything about how one ought to respond to the actions of these agents.

In Chapter One, I listed a range of possible responses to agential behaviour. This list was roughly ordered from responses that had little to no effect on the agents to whom the response is directed, to those that had direct and intentional effects; the responses ranged from mere moral evaluation, through emotional expressions of approval and disapproval, to certain types of punishment and reward.

I do not intend to claim that flexible moral agents are fitting targets for responses further down in the list. Being able to guide one's behaviour by moral reasons does not necessarily warrant punishment in cases where agents fail to properly guide their behaviour, for instance.[41]

Rather, the appropriate response to the behaviour of flexible moral agents, in virtue of their sophisticated abilities, is a *normative expectation* that such agents will use these abilities to their potential. If one can justify one's behaviour, then one ought to ensure one's behaviour is justifiable. And if one can improve one's moral character, then one ought to try to become better.

Specifically, I take these to be *moral duties* of flexible moral agents. Of course, one can reasonably question whether this is in fact the case, and if so, how these duties relate to the moral foundations of welfare, autonomy, and fairness. Here is the argument. Moral agents ought to respond appropriately to considerations of welfare, autonomy, and fairness. But this ability is not completely effective. At times, moral agents fail to respond appropriately to these considerations; agents may be insensitive to these considerations,

---

the question of where these desires derive their content. I hope to have offered a plausible suggestion in this section.

[41] Although Fischer (2004), as we saw in Chapter Two, appears to disagree on this point.

or suffer from weakness of will, or have a stronger desire to do something other than respond appropriately to these considerations. At these times, moral agents fail to do the right thing.

This implies that moral agents ought to reduce the number of times they fail to respond to considerations of welfare, autonomy, and fairness, insofar as it is possible for them to do so. For flexible moral agents, two ways of doing this are ensuring that their actions are morally justifiable and improving their moral character. By ensuring that their actions are justifiable, agents commit to responding appropriately to the relevant moral considerations. And by improving their moral character, agents improve their ability to respond appropriately to these considerations. In both cases, agents fulfil a moral duty to act rightly more often. As far as moral duties go, this is relatively uncontroversial and neutral with respect to theories of normative ethics.

Because basic moral agents lack these abilities, we cannot hold them to the same standards. In blaming such an agent, we make a judgement of their moral character but we do not expect better of them. If we want a basic moral agent to act better, we cannot appeal to their reason, as only flexible moral agents are capable of using moral reasons. We can use nonrational methods, such as behavioural conditioning, but as we have seen, this will only result in reliable moral improvement if the person doing the conditioning is herself using moral reasons.

Conclusion to Chapter Three

At the beginning of the thesis, I gave a schema of a definition of moral agency, according to which moral agents are those agents who have certain abilities, which enable the performance of certain actions. Because moral agents have these abilities, it is appropriate to respond to them in certain ways.

In the first two chapters, I developed an account of *basic moral agency*, in which I filled out this schema by claiming that basic moral agents are able to have desires about others' mental states, which enables them to be motivated by moral reasons, and because of this, it is appropriate to evaluate them as praiseworthy or blameworthy in virtue of their behaviour.

In this chapter, I have given a more restricted account – *flexible moral agency* – which fills out the schema in a different way. In particular, it aims to make sense of the more flexible moral behaviour of agents such as ourselves. According to this account, flexible moral agents are able to have and be guided by moral reasons, which enables them to justify their actions and improve their moral character, and because of this, it is appropriate to expect and to help them to do so.

Much of this chapter has focussed on spelling out the differences between moral guidance, characteristic of flexible moral agency, and the simpler ability of moral motivation, characteristic of basic moral agency. I have argued that while moral motivation need only involve desires about others' mental states, moral guidance involves comparative beliefs about desires about others' mental states.

Having established some demarcation criteria for basic and flexible moral agency, we may wonder how these abilities could arise in the first place. How is it possible for a world to be completely devoid of moral agency, as presumably was the case at some earlier point in our evolutionary history, and at some later point to be populated with moral agents? I offer one type of answer in the next chapter. Unlike much of the existing work on this

topic, I do not aim to offer an account of the evolution of morality.[42] While this is valuable work, its wide appeal across a range of disciplines leaves me with little to add. Instead, I answer this question by considering our ability to use mental representations. Specifically, I consider changes in children's ability to represent and to understand representation over the first four years of their lives, as well as evidence from other animals that correspond to particular stages of childhood development. I will use this framework to identify when both basic and flexible moral agency develop in early childhood and whether any nonhuman animals are moral agents of either the basic of flexible kinds.

[42] See, for example, Sober & Wilson (1998), Richerson & Boyd (2005), Joyce (2005).

CHAPTER FOUR: THE DEVELOPMENT OF MORAL AGENCY

In the first two chapters, I offered an account of moral agency in which moral agents are characterised by their ability to be motivated by moral reasons, and I argued that this was constituted by the ability to have desires with moral content; that is, desires about others' mental states. In virtue of their ability to have these desires, it is appropriate to evaluate moral agents as praiseworthy for performing morally right actions and blameworthy for performing morally wrong actions, assuming that they are able to act upon these desires. I called this kind of moral agency "basic moral agency".

In the previous chapter, I outlined a more restrictive account in which more sophisticated moral agents have the additional ability to be guided by moral reasons, where this is understood as the ability to use moral judgements. I argued that this ability enables these moral agents to justify their actions and improve their moral character. In virtue of this, I claimed that these agents have a responsibility to ensure their behaviour is justifiable and to improve their moral character. I called this kind of moral agency "flexible moral agency", as the ability to be guided by moral reasons presupposes and builds upon the ability to be motivated by moral reasons, and enables more flexible moral behaviour.

I also observed that both types of moral agency involve the use of mental representations. For basic moral agents, moral motivation involves mental representations – desires – with a specific kind of *representational content*: others' mental states. For flexible moral agents, moral judgement additionally involves different mental representations – beliefs – with a different kind of content: the desirability of certain actions.

As children develop, they become more proficient in the use of mental representations. In particular, research on *theory of mind*, the ability to attribute mental states to others, has shown that children develop particular representational abilities at particular ages. For instance, children almost invariably develop the ability to attribute false beliefs to other people before their fifth birthday.[1] Given that both basic and flexible moral agency involve the specialised use of mental representations, it is worth considering how theory of mind research bears on the development of moral agency, and in particular, whether it can tell us the age at which children develop both basic and flexible moral agency.

In the 1980s, the developmental psychologist Josef Perner developed the theoretical framework in which much of this research has been conducted, and which has been well-supported by the empirical evidence over the last forty years.[2] This framework distinguishes between three increasingly complex representational abilities, which develop sequentially in children: *primary representation* is characterised by the ability to represent one's immediate environment, *secondary representation* is characterised by additional representations 'decoupled' from reality, and *metarepresentation* is characterised by the ability to recognise representational properties. It is my contention that basic moral agency requires secondary representation but not metarepresentation, whereas flexible moral agency requires metarepresentation.

---

[1] Wellman *et al* (2001)

[2] This framework is most extensively developed in Perner (1991). Subsequent research, including a meta-analysis by Wellman *et al* (2001) and review by Doherty (2007), has largely corroborated the framework's applicability to childhood development, and it has been extended to nonhuman animals by Whiten & Suddendorf (2001). I consider critiques in Sections Two and Four.

In the following section I give an overview of primary, secondary, and metarepresentation. I focus specifically on the distinction between secondary and metarepresentation and why the former is necessary but not sufficient for the latter.

In the second section, I observe that while this framework is justified by the empirical evidence, it relies on a misunderstanding of a distinction between two different functions of representation, and because of this, it has trouble accommodating desires. I propose a modification to the framework, which aims to accommodate desires while remaining consistent with the empirical evidence.

In the third section, I situate basic and flexible moral agency within this framework. I consider various types of morally relevant mental states, including beliefs and desires, phenomenally conscious states, and reactive attitudes. I argue that agents cannot be motivated by desires about these mental states unless they are capable of secondary representation and that there are at least some cases of moral motivation for which metarepresentation is not necessary. I also argue that moral judgement requires the ability to conceive of desires as representational, and thereby requires metarepresentation.

In the final section, I give an overview of the empirical evidence for the development of secondary and metarepresentation in children, and for their presence in nonhuman animals. I primarily consider two abilities indicative of secondary representation: pretend play and mirror self-recognition; and two abilities indicative of metarepresentation: the ability to attribute false beliefs and the ability to inhibit one's actions. On the basis of this evidence, I tentatively conclude that children typically develop basic moral agency from around 18 months and flexible moral agency from around 3.5 years, and that basic moral agency may be present in a few nonhuman species but that flexible moral agency is not.

A Framework for the Development of Metarepresentation

In this section I give an overview of Josef Perner's framework for the development of metarepresentation in children. The focus of this section is describing each of the three stages – primary representation, secondary representation, and metarepresentation – and how they relate to each other. I begin with some preliminaries about representation in general, and describe two different functions of representation, which form the basis of the distinction between primary and secondary representation. I then discuss Perner's arguments for the claim that secondary representation is necessary but not sufficient for metarepresentation.

In thinking about representation, it helps to first consider cases of *non*-mental representation, such as pictures and words. Consider a painting of a tree. Here, we can distinguish between three things that are relevant to this painting as a representation: the *representational medium* is the painting as a physical object – the arrangement of paint on the canvas; the *represented object*[3] is the real tree; and the *representational content* is the tree as depicted in the painting.[4] We can make the same set of distinctions with respect to the written word 'tree': the medium is the arrangement of letters on the page, the object is the tree to which the word refers, and the content is that tree as imagined by the reader.

[3] This is not to say that all represented objects are in fact objects. We can have representations of nonexistent objects, such as unicorns, and of entities that are not objects, such as the French Revolution. I shall use the term 'represented object' regardless of whether the entity in question is a real object, or an object at all.

[4] Crane (2003)

Mental representations, such as beliefs and desires, also involve this relationship between medium, object, and content. In the case of both beliefs and desires, the medium is the mental state itself. Just as a tree can be represented by a word or a picture, it can be represented by a belief or a desire. Strictly speaking, however, beliefs and desires do not simply represent objects such as trees; they represent *situations*.[5] For instance, my belief that there is a tree in front of me does not simply represent a tree, but the situation in which there is a tree in front of me. This is also true of desires, though it is a little less obvious. Although I might say that I desire a coffee, this can be better described as my desiring *that* I have a coffee. Thus, the desire is about the situation in which I have a coffee.

As with words and pictures, the representational content of the relevant belief or desire is the situation *as represented by the representational medium* – in this case, the relevant belief or desire, – and the represented object is the *actual situation*. Two points are worth noting here. First, the actual situation may differ from the represented situation, as is the case for false beliefs. If I believe that there is a tree in front of me but the object in front of me is actually a sculpture, then there is a mismatch between the content and object, and my belief *misrepresents* reality. Second, the actual situation may be purely hypothetical. This is generally the case for desires, which do not aim to represent reality as it is, but to motivate agents to bring about the represented situation. My desire for a coffee does not represent a (currently) real situation but it does motivates me get a coffee, such that the situation comes into alignment with the content of the desire.[6]

I will focus on beliefs here and I will consider desires in the following section. This is because the research on theory of mind has largely focused on beliefs, due to the fact that desires cannot misrepresent.[7] Given this, psychologists have generally attempted to study theory of mind and metarepresentation by testing whether children understand that beliefs can be false.[8] I will discuss this research in more detail in Section Four.

In this section, I will discuss a different distinction between types of beliefs. Roughly speaking, we can distinguish between beliefs that aim to represent reality and those that aim to represent hypothetical scenarios.[9] An example of the former is the belief that I am standing under direct sunlight. An example of the latter is the belief that I would be more comfortable if I were to stand under a tree. This is the distinction between beliefs that use *primary representation* and those that use *secondary representation*. Perner uses these terms because he takes the former to be prior to and necessary for the latter.[10]

He illustrates this with several examples of nonmental representation, including maps and sandbox models. For instance, the purpose of a map is to allow people to find their way, given the regularity between symbols on the map and locations in the world. However, once we understand the concept of a map, they can take on secondary functions. We can create maps of fictional environments or of projected changes to the world.[11]

This ability to have beliefs about hypothetical scenarios is a particularly useful one to have. Perner illustrates this using sandbox models of a battlefield. A primary model serves the

[5] Crane (2003)
[6] This is sometimes referred to as world-to-mind direction of fit, as opposed to the mind-to-word direction of fit of beliefs: see, for instance, Anscombe (1957) and Searle (1983), but see Sobel & Copp (2001) for the opposing view. I will avoid this terminology, as it is not obviously true of beliefs involving secondary representation, and will needlessly complicate the discussion of those beliefs.
[7] Dennett (1978)
[8] Wellman *et al* (2001)
[9] Perner (1991)
[10] *Ibid.*
[11] *Ibid.*, pp. 24-25

primary function of representation by providing accurate information about the environment. This would be useful during battle, as generals could use it to know the locations of their own and opposing armies, but they would be better served by the addition of extra models to serve the secondary function of representing hypothetical states of affairs, which could be used to develop and test plans of attack.[12] A major benefit of the secondary function of representation, then, is that it allows greater flexibility in behaviour, by allowing one to generate and test plans of action.

The notion of models plays an important role in Perner's framework. Just as a physical sandbox model represents a situation, so too does a mental model. One could have a mental model of the real battlefield and a separate mental model of a hypothetical battlefield. It is important to note the relationship between models and representations, particularly in the context of discrete mental representations, such as beliefs and desires. On this picture, a model is a set of representations that collectively represent a particular situation. For instance, I have a 'primary' model of the world, comprised of all my beliefs about the way the world is. I also have various 'secondary' models, each comprised of sets of beliefs about the way the world could be, given certain counterfactual scenarios. For the sake of clarity, I will use the term *mental representation* to refer to discrete beliefs and desires, and *mental model* to refer to sets of mental representations that serve to represent a particular situation.

Strictly speaking, secondary representation need not represent hypothetical scenarios. It is more accurate to say that it involves representation (models) of scenarios that *appear* to differ from reality. This includes models of the same scenario from different perspectives, as is the case when we recognise our reflection in a mirror. The theory thereby predicts that children develop this ability at around the same time as they develop the ability to consider hypothetical situations. This is borne out by the evidence, which I discuss in Section Four, that children begin both to recognise their reflection and to engage in pretend play at one to two years of age.

Perner's characterisation of secondary representation, then, is one in which representations gain additional functions beyond simply representing a single model of reality, including both hypothetical nonreal models and models of reality from other perspectives. This ability to represent multiple models of reality is necessary for the third stage, metarepresentation.[13] To see this, it is important to first understand that metarepresentation is not simply representation of representation. If a pre-literate child were to look at the written word "tree", she would have a visual representation of this word, which itself is a representation of a tree. But this is not metarepresentation because the child does not understand the connection between the word and the object. Genuine metarepresentation is characterised by *representation of the representing relation*: in this case, representation of the meaning of the word.[14] More generally, the representing relation is the relation between the representational content, the representational medium, and the represented object. This requires multiple models of reality, for instance representations of the word "tree" not merely as a series of marks on a page but also as a word with a specific meaning, as a representational medium with specific content. Moreover, multiple models are required to make sense of the relation between this content and the represented object. For instance, if a shrub is mistakenly described as a tree, one can understand this as a case of misrepresentation only if one represents the object, the misdescribed shrub, as being different from the content, the imagined tree.

[12] Perner (1991), Dennett (1996)
[13] Perner (1991)
[14] *Ibid.*

But while multiple models are necessary for metarepresentation, they are not sufficient. To show this, Perner discusses an ability that is almost but not quite metarepresentation: drawing inferences from correspondences between multiple models. At first, this may seem equivalent to metarepresentation. For instance, a pre-literate child can identify a picture of a tree as a tree, based on the similarities between the picture and real trees. Perner claims that this falls short of genuine metarepresentation because although the child makes an inference about the picture, this inference is not an *interpretation*.[15]

One can make sense of this claim by considering cases of misrepresentation. Suppose that the picture is labelled as a tree but instead depicts a shrub. If I infer that the picture depicts a tree, and I then discover that it in fact depicts a shrub, this does not on its own show that I interpret the picture as misrepresenting the shrub. I may treat the picture *as if* it is a misrepresentation, but *interpreting* the picture as such requires understanding that the picture is *supposed* to represent a shrub. In other words, the representing relation between the content and the object is that the former has the *function* of representing the latter. Without the ability to represent this function, one cannot conceive of the function being unfulfilled; one cannot conceive of the picture *mis*representing the shrub.[16] Thus, to make this or any other interpretation of a representation, one must not only represent the content and the object, which requires secondary representation, but also that the function of the content is to represent the object, which goes beyond secondary representation.[17]

Recall that secondary representation is characterised by the presence of multiple models, which are distinguished from the primary model and from each other by their function. A toddler may have a primary model of reality, as well as secondary models that represent hypothetical scenarios (such as in pretend play) and reality from other perspectives (such as in mirror self-recognition). The fact that toddlers do not mistake their pretend play for reality or their reflection for another child indicates that these multiple models are easily distinguished from one another, as though they are 'tagged' with their respective functions. Metarepresentation seems to add an additional tag: that of representation itself. When I read the word "tree", the visual representation of the word on the page is accompanied by another representation of a tree. The relation between the two is that the former represents the latter. This is a different relation than that between real and hypothetical scenarios, or between the same scenario from different perspectives. The ability to represent this representational relation is what distinguishes metarepresentation from mere secondary representation.

This is Perner's account, which uses belief as a paradigm example of representation to distinguish between the three levels of primary representation, secondary representation, and metarepresentation. As mentioned above, desires pose a problem for this account because they have a different function but are equally necessary for intentional action. In the following section, I aim to situate desires within this account, modifying it as necessary to do so.

The Place of Desires within the Framework

Perhaps because the research on the development of representation has largely focused on theory of mind, and in particular, children's ability to understand misrepresentation by attributing false beliefs to others, desires have been relatively neglected in this research. Not wholly neglected, particularly in the literature on executive function – the ability to

[15] *Ibid.*
[16] Dennett (1978), Perner (1991)
[17] Perner (1991)

inhibit or control one's behaviour, – but it is clear that desires do not fit neatly into Perner's developmental framework.[18]

One the one hand, desires seem like a paradigm example of secondary representation. They represent goal states that are typically different from current reality. In this way, the desire for a cup of coffee, for instance, is like the 'secondary' belief that it would be good to have a coffee and unlike the 'primary' belief that I am currently drinking a cup of coffee. On the other hand, beliefs and desires seem interdependently necessary for intentional action.[19] My action of going to the café can be explained by reference to a relevant belief-desire pair: the belief that the café serves coffee and the desire that I have a coffee. Without either, I would not have gone to the café.

This poses a problem for Perner's account, as he takes primary representation to be conceptually and developmentally prior to secondary representation.[20] This would seem to have the consequence that desires require beliefs but not vice-versa, despite the fact that both seem equally necessary for intentional action.

Moreover, just as Perner uses examples of nonmental representation, such as maps and sandbox models, to distinguish between the primary and secondary functions of representation, we can use other nonmental examples to show that desire-like functions of representation need not depend on belief-like functions. For instance, traffic lights have the desire-like function of getting drivers to stop or go, depending on the colour. Unlike map reading, which depends on a *prior correlation* between the cartographic symbols and locations in the real world, obeying a traffic light *constitutes the correlation* between its colour and the movement of traffic. And unlike a map, in which the symbols bear the same spatial relationship to one another as their real world counterparts, there is no obvious corresponding relationship between colour and movement in the real world, except insofar as this relationship is created by drivers obeying the lights. This suggests that traffic lights do not derive their desire-like function of motivating drivers to stop and go from any prior belief-like function, which in turn suggests that desires serve a primary function, not a secondary one.[21]

The only way forward, it seems to me, is to reject Perner's framework, either in whole or in part. As we shall see in Section Four, there is substantial empirical evidence in favour of this framework, so it would be unfortunate to reject the whole thing, but perhaps we can modify part of it in such a way that it better accommodates desires while remaining consistent with the empirical evidence. The most sensible part to reject is the claim that representations have primary and secondary *functions*, since the issue arises from the fact that desires do not fit neatly into either category.

---

[18] See, for instance, Schwitzgebel (1999)

[19] Hume (1738/1985), Dretske (1991), Crane (2003)

[20] Perner (1991), Schwitzgebel (1999)

[21] Eric Schwitzgebel (1999) also claims that Perner's framework has trouble accommodating desire. He points to a contradiction between the claim that for any representation it is possible to misrepresent, and the claim that desires, which cannot misrepresent, are nonetheless representations, both of which Perner seems to accept (Perner 1991, pp. 20, 116, 144, 205; quoted in Schwitzgebel 1999, pp. 162-164). Schwitzgebel attributes this error to the failure of philosophers, whose accounts form the basis of Perner's, to distinguish between *contentive* accounts of representation, in which representations are representations in virtue of having content, and *indicative* accounts, in which this content is supposed to match up with the way things are in the world (1999, pp. 158-159). My dispute is different. I take for granted that representation need not be indicative, and take Perner to be saying the same thing by positing a nonindicative secondary function for representation. Rather, I argue that this nonindicative function, at least in the case of desires, is not derived from the "primary" function of indicative beliefs.

However, to reject this claim would still entail a major change to the framework, since the developmental changes associated with secondary representation, such as pretend play and mirror self-recognition, are well explained by the ability to use representation for secondary functions. Given that secondary representation is a unifying explanation for these changes, it is worth exploring whether we can reconstruct the concept of secondary *representation* without relying on the concept of secondary *functions*.

One way of approaching the issue is to ask what, other than function, distinguishes primary from secondary representation. One feature seems obvious: *number of models*. As we have seen, secondary representation is characterised by the ability to use multiple models. However, this seems like little more than a redescription of secondary functions, as these extra models are distinct from the primary model in virtue of their secondary functions. Moreover, desires also seem to involve additional models, given that they typically represent goal states that differ from one's primary model of reality. For instance, if I desire a coffee and believe that I do not currently have one, then the content of this desire seems to belong not to my primary model, but to a different one.

A more promising distinguishing feature is *content availability*. By this, I mean that representational content from one model is available for use by another model. This is consistent with what we already know about secondary representation. Mirror self-recognition, for instance, appears to involve integration of content from a secondary model with that of the primary model. Consider a standard test for mirror self-recognition, the mark test. In this test, which has been administered to both young children and nonhuman animals, subjects are exposed to mirrors in the initial familiarisation phase (or begin the experiment already familiar with mirrors) and are later marked imperceptibly on their eyebrow or some other part of their body that is only visually accessible using the mirror. Subjects are deemed to have passed the test if there is an increase in touching or other behaviour directed at the mark, compared with the period prior to marking, after looking in the mirror.[22]

Now consider the operation of the primary and secondary models. The primary model includes the subject's visual reflection, as well as any tactile representation of the mark. This raises the question of how the subject can infer from their reflection where to touch. The answer, for subjects capable of secondary representation, is by using content from a secondary model.[23] This model includes a representation of the subject's own face and the mark on it, from an external perspective. It is this external perspective that marks the model as secondary rather than primary. The content from this secondary model is then used by the subject to direct their touch, and in so doing, to add additional information to the primary model. Content availability is the ability for agents to use information from one model in this way to update information in another model. Prior to the development of secondary representation, it seems that content from other models cannot be used in this way,[24] and this inability explains why younger children and most animals tend to fail the mirror self-recognition test.

Now consider desires. If secondary representation is distinguished by content availability, then this implies that content from desires could be integrated into the primary model after and only after the development of secondary representation. It is important to note that this integration is more complex than mere intentional action. At first glance, it certainly seems like intentional action involves this kind of content integration. After all,

[22] These experiments are discussed in more detail in Section Four. The canonical experiment was performed by Gordon Gallup (1970) on chimpanzees.
[23] Asendorpf *et al* (1996), Nielsen & Dissanayake (2004)
[24] Perner (1991)

desires aim to motivate agents so as to make their content *become reality*, which may then provide updated information to the primary model. My desire for a coffee represents the scenario in which I have a coffee, it motivates me to get a coffee, and thereby creates a reality in which the content of the desire matches reality. At that point, I might then form the belief that I now have a coffee, thereby integrating the content from the desire *indirectly* into my updated primary model of the real world. An important difference is that in the case of mirror self-recognition, the secondary model provides updated information to the primary model *directly*, without changing reality first, whereas this is not the case for the desires we have just discussed.

But desires can change the primary model directly, unmediated by action, and the empirical evidence suggests that they begin to do so at around the same time as other indicators of secondary representation emerge, such as mirror self-recognition and pretend play. One example, as observed by Perner, is that of emotional meltdowns. These typically begin between a child's first and second birthday, and subside considerably before their fourth, thereby fitting within the developmental timeframe of secondary representation.[25]

Perner observes that young infants seem not to experience frustration. Instead, they tend to persist with a goal, such as reaching for an object, until they either succeed or become distracted. Note that this is not to say that infants do not experience *distress*, such as that caused by pain or hunger, but rather that they do not experience the specific kind of distress caused by unfulfilled desires. Toddlers, by contrast, seem to experience frustration, in the form of meltdowns, when their desires are unfulfilled. It is as if toddlers, but not infants, are *aware* of their unfulfilled desires.[26]

This provides additional evidence for increased content availability beginning at between one and two years of age. In the case of a toddler trying to reach for an object, their primary model represents the object as being out of reach, and, as informed by their desire, as being an unpleasant situation. The separate lines of evidence from mirror self-recognition, pretend play, and emotional meltdowns, all suggest that these newfound abilities are due to improved content availability. Specifically, that the primary model can now integrate representational content from other models, including beliefs about hypothetical situations, beliefs about situations from other perspectives, and desires.

What about content availability in other directions? Could the primary model influence secondary models? Or could secondary models influence other secondary models? And is there evidence suggesting that this emerges at around the same time as content availability influencing the primary model? This is especially relevant for secondary models comprised of *desire content*, since moral motivation is characterised by a specific kind of desire content, namely content about others' mental states.

One line of evidence suggesting that this can occur is the emergence of a new type of desire in toddlers: the desire to do things for oneself. It is a well-documented fact that toddlers, unlike infants, often want to do things for themselves and will often resist help from others.[27] For instance, when building a tower of blocks, infants and toddlers seem to be motivated by distinctly different desires. The infant seems to desire merely that the tower be built, while the toddler seems to desire that the tower be built *by oneself*. Assistance from a parent fulfils the desire in the case of the infant but frustrates the desire in the case of the toddler, thereby explaining their different reactions. This development seems readily explainable by the changes in content availability associated with secondary representation.

[25] Perner (1991), Sroufe (1997), Lieberman (2017)
[26] Perner (1991)
[27] Geppert & Küster (1983), cited in Perner (1991, pp. 221-222). See also Moore (2010), p. 43.

Here, it seems that a new type of representational content – doing things for oneself – is made available to desires from the another model.

Consider that toddlers' resistance to help shows that they have a *belief* with specific content along the lines of "the tower is not being built by myself", which shows that the "doing it by myself" content is present in beliefs. These beliefs may supply desires with this kind of content. Now, it is not obvious to me which *model* supplies desires with this content, whether it be the primary model or a secondary one. On the one hand, if *infants* can distinguish between their own actions and those of others, then this would suggest that the content comes from the primary model. On the other, if the concept of *oneself* requires more sophisticated abilities, such as mirror self-recognition, then this suggests that a secondary model is needed to supply this content. However, while this remains an open question, we do not need an answer to see that this new content could be supplied by some other model, a model constituted by the content of beliefs.

This is supported by various accounts of the generation of new desires – specifically, new *intrinsic* desires. Intrinsic desires are those that represent situations that are wanted for no further reason. By contrast, *instrumental* desires represent situations that are wanted as a means to something else.[28] The desire for happiness and the desire for money are paradigm examples of intrinsic and instrumental desires, respectively. The generation of new instrumental desires is not difficult to explain: if I intrinsically desire happiness and come to believe that money will make me happy then I will come to instrumentally desire money. But the generation of new intrinsic desires is more difficult to explain and there is disagreement over how this occurs.[29]

Michael Smith, along with his co-authors, has claimed that when agents deliberate rationally, their moral judgements can cause them to form the corresponding desire.[30] For instance, if an agent deliberates and judges that she ought to donate to charity, then all else being equal, she will desire to donate to charity. This is not to suggest that *toddlers* are capable of deliberation or of making moral judgements, but it's not difficult to see how secondary representation could be implicated in the generation of new desires from such judgements. The agent who judges that she ought to donate to charity has a representational model of the situation in which she donates to charity. Because she is capable of secondary representation, this content is available for use in other representational models. In this case, it seems she has a model of what she *ought* to do, constituted by the content of her moral judgements, which supplies content to a model of what she *desires* to do, constituted by the content of her desires. In this way, moral judgements can generate new desires by providing representational content from moral judgements to new desires.

Neil Sinhababu, unlike Smith, denies that deliberation can ever produce new intrinsic desires. However, he does recognise that intrinsic desires may be generated by nondeliberative processes, such as via conditioning.[31] Timothy Schroeder likewise offers an account in which conditioning may generate new intrinsic desires.[32] Again, it is not difficult to see how secondary representation could facilitate this. It seems plausible that children have an innate intrinsic desire for novelty, and that the transformation of a pile of blocks into a tower could fulfil this desire. By repeatedly building block towers, this transformation could become associated with the process of doing things for oneself, and

[28] Schroeder (2020)
[29] *Ibid.*
[30] See, for example, Smith (1994), Kennett and Smith (1996), Pettit and Smith (1990).
[31] Sinhababu (2017), p. 4
[32] Schroeder (2004)

toddlers could come to intrinsically desire this. It seems plausible that repeated associations between the content of different representational models effectively 'copies' content from one model to another. In this case, there seems to be a belief-like model that represents (a) novelty and (b) doing things for oneself, and a desire-like model that represents (c) novelty, and the repeated association of these elements causes (b) to be copied to the desire-like model, resulting in a new desire to do things for oneself.

A potential objection to this line of reasoning is that conditioning does not seem to require secondary representation. For instance, Schroeder gives the example of an infant who intrinsically desires food, warmth, and human contact, and comes to intrinsically desire her mother's presence due to the repeated association of her mother with these things.[33] Insofar as the latter desire seems to develop later than the former desires, it strikes me that this could be equally well explained by a pre-existing *intrinsic* desire for her mother's presence that takes a little longer to become apparent due to the relatively slow development of vision in infancy, compared with touch.[34]

Taken together, the lines of evidence provided by other mirror self-recognition, pretend play, emotional meltdowns, and the desire to do things for oneself, particularly given that they all develop at between 12 to 18 months of age, independently suggest that content availability between models is a defining characteristic of secondary representation.

What about metarepresentation? Do the above changes to the concept of secondary representation entail any changes to the concept of metarepresentation? I think not. There is nothing in Perner's original analysis of metarepresentation that poses a problem for desires in the way that the distinction between primary and secondary functions pose such a problem. Recall that metarepresentation is characterised by the ability to represent the relations between representational content, representational media, and represented objects, or more prosaically, *representation of representations as representations*. Although Perner and many other theory of mind researchers have focussed on beliefs about beliefs, his analysis of metarepresentation also allows for the possibility of other combinations, including beliefs about desires, desires about beliefs, and desires about desires.[35]

One issue still requiring clarification is the relationship between secondary representation and metarepresentation. Recall that in Perner's original analysis, secondary representation was necessary but not sufficient for metarepresentation. This is also true of my revised analysis, because the difference between the two analyses is small. Although I did reconceptualise secondary representation as necessarily involving content availability, this was already implicit in Perner's framework, as shown by the above discussion of mirror self-recognition and other examples, all of which are drawn from Perner's own work. Rather, the only major difference between Perner's analysis and my own is that I have done away with the distinction between *primary and secondary functions* of representation. This, however, does not entail the elimination of *representational functions* altogether, only the claim that a single one of these functions is necessary for all other functions. As we have seen, individuals use several representational models with distinct functions, including the following:

1. a 'primary' model that functions to represent the world as it is,
2. a model of one's goals, as represented by one's desires,

[33] *Ibid.*
[34] Bornstein *et al* (2013)
[35] Perner even references Frankfurt's discussion of second-order desires (that is, desires about desires), although he does not discuss these in detail: Frankfurt (1971), referenced in Perner (1991).

3. models of hypothetical or nonreal scenarios, involving counterfactual beliefs,
4. models of real scenarios from other perspectives, and
5. models of representations.

Of these types of models, (1) and (2) seem to be present from infancy and are associated with primary representation.[36] Developing between the ages of one and two years are models of types (3) and (4), and while Perner takes their presence to be indicative of secondary representation, I claim that the crucial development at this stage is the general ability for representational content to be shared between different models. Perner notes that metarepresentation – models of type (5) – depend on the existence of multiple other models, such as a 'primary' model of the representational medium (say, a picture of a tree) and a 'secondary' model of a represented object (the real tree). It is implicit in this framework that content from the various models is available for use by other models, thereby making it possible to compare the picture with the tree and to identify, for instance, whether the picture accurately represents the tree or whether it misrepresents the tree. Thus, secondary representation is still necessary for metarepresentation.

Similarly, secondary representation is still insufficient for metarepresentation. The presence of multiple representational models that can draw upon content from one another does not entail anything about the various *functions* of these multiple models. Specifically, it does not entail that any of the models function to represent representations.

We now have a coherent and empirically-supported framework in hand. It is summarised for clarity in the table below:

Primary Representation
Present during early infancy and characterised by self-contained representational models. These include a 'primary' model of the world, constituted by content from beliefs about the way the worlds is, and additional models of goal states, constituted by content from desires.
Secondary Representation
Develops at between one and two years of age and characterised by multiple models differentiated by function, which can share content with one another. Additional functions include representing hypothetical nonreal scenarios and representing real scenarios from other perspectives. Associated with several developmental changes, including mirror self-recognition, pretend play, emotional meltdowns, and wanting to do things for oneself.
Metarepresentation
Develops at between three and four years of age and characterised by the ability to represent representations as representations. Associated with several developmental changes, including the ability to attribute beliefs to others and to inhibit one's actions (discussed in more detail in Section Four).

We are now in a position to consider the relationship of this framework to moral agency. As I shall argue in the following section, *moral motivation*, and thereby basic moral agency,

[36] Bornstein *et al* (2013), Perner (1991)

requires secondary representation but not metarepresentation, whereas *moral judgement*, and thereby flexible moral agency, requires metarepresentation.

Representation of Others' Mental States

In this section, I will consider the representational abilities required for moral motivation and moral judgement. Both involve representation, as the former requires a specific desire, while the latter requires a specific belief. As previously argued, moral motivation involves a desire about another's mental state, whereas moral judgement involves a belief about the desirability of an action. The discussion will primarily focus on moral motivation, although in the course of this discussion, the representational abilities for moral judgement will become clear.

Recall from Chapter One that acting morally involves responding to others in one of three ways, out of concern for their welfare, or for their autonomy, or for fairness; and in each case, to be motivated by these concerns is to act on a desire about another's mental state. In order to determine the representational abilities involved in moral motivation, it is important to identify and examine the relevant mental states to which moral agents respond.

Consider first acting out of concern for others' welfare; acting so as to prevent their suffering and to promote their best interests. One prominent theory of welfare is hedonism, according to which welfare consists in *experiential states*, such as pleasure, pain, happiness, and so on.[37] A second prominent theory is the preference-satisfaction account, according to which welfare consists in the satisfaction of preferences. These are may conceived of as satisfied *desires*, or as *beliefs* about what is good.[38] The third major type of theory is that of objective list theories, which have a pluralist conception of welfare. These theories typically include things in addition to experiential states and satisfied preferences, but the most common additions are typically either mental states or reliant on knowledge of mental states.[39] For instance, objective list theories often include knowledge and relationships with others as items on the list. Knowledge is generally considered a type of *belief*; to know something is to believe it.[40] Relationships are not a mental state but it seems that they necessarily involve the communication of *reactive attitudes*, such as gratitude and resentment,[41] which are themselves mental states.

Now consider acting out of respect for others' autonomy; acting in such a way that other agents are free to pursue their goals without interference. There is debate over which agents have the relevant autonomy[42] but it is clear that autonomous agents are capable of intentional action, because without this ability, they cannot pursue goals or even act at all. Since intentional action requires a belief-desire pair, agents respect each other's autonomy by ensuring that the other's *beliefs* and *desires* are not interfered with.

Finally, consider acting to ensure fairness. Since fairness involves the just distribution of desirable goods, one must be able to conceive of a good as being desirable – as being the

[37] Crisp (2006)
[38] Singer (1979)
[39] Griffin (1986)
[40] Ichikawa & Steup (2017)
[41] Strawson (1962) offers the canonical argument for the necessity of the reactive attitudes to relationships.
[42] The Kant scholars Wood (2007) and Korsgaard (2017) have both argued that animals have some limited autonomy, against the standard Kantian claim that this is not the case.

type of thing that ought to be the *content of a desire* – in order to judge whether it is distributed fairly or unfairly.

So the relevant mental states include beliefs, desires, experiential states, and reactive attitudes. Given that motivation by moral reasons involves desires about any of these mental states, if any are not themselves representations, then this would show that moral motivation does not require metarepresentation. Since beliefs and desires are representations, we can concentrate our initial investigation on experiential states and reactive attitudes.

A common way of categorising mental states is to distinguish between *psychological* states and *phenomenal* states. This distinction, which has its roots in philosophy of mind,[43] characterises psychological states as essentially *functional* whereas phenomenal states are essentially *experiential*. Beliefs and desires, which function to produce intentional behaviour are examples of psychological states, while experiential states, such as pain or the visual experience of seeing the colour red, are examples of phenomenal states.

It may be noted that mental states frequently seem like they are both psychological and phenomenal. Desires, for instance, motivate agents to act but they also seem to feel a certain way. If I desire a coffee, then I typically experience a 'pull' toward the café. Similarly, experiential states often seem to serve important functions. If I were to feel pain upon touching a hot stove, for instance, this would not only feel uncomfortable but it would cause me to pull my hand away from the source of the heat. It seems, then, that many, maybe all mental states are both psychological and phenomenal.

Perhaps a better way of drawing the distinction is to speak of psychological and phenomenal *aspects* of mental states. It seems to me, though, that even if these aspects do not in fact pick out different types of mental states, they are *conceptually* distinct, such that one could coherently imagine mental states that are either psychological or phenomenal but not both. This is the thought behind *philosophical zombies*, which are physically identical to humans in every respect, but who lack phenomenal mental states entirely.[44] In the language of philosophers working in the area, there is *nothing it is like to be* a philosophical zombie. It strikes me that we could explain the behaviour of zombies in much the same way as we describe the behaviour of humans, by reference to their beliefs and desires. The difference would be that they wouldn't feel the pull of their desires as we do. It wouldn't feel like anything to the zombie to have these beliefs and desires.

The claim that zombies are even possible is controversial, given that a physically identical copy of a human seems like it should be identical in all respects, including its phenomenal mental states.[45] But we can imagine more prosaic examples. When I scan a bottle of milk at the supermarket checkout, the self-checkout machine uses an optical sensor to detect the barcode and uses this information to charge me the predetermined amount of money for a bottle of milk. It is using psychological states to perform this behaviour, in just the same way as a human checkout assistant would do so in order to charge me the correct price. But unlike the human checkout assistant, it doesn't seem to me that the self-

[43] I take this terminology from David Chalmers (1996). Ned Block (1995) draws a similar distinction between access consciousness and phenomenal consciousness.
[44] Kirk (1974), Chalmers (1996)
[45] The literature here is extensive, but the Stanford Encyclopedia of Philosophy article on zombies (Kirk (2021)) gives a good overview.

checkout machine has a *visual experience* of the bottle of milk. It seems to have psychological mental states in the absence of phenomenal mental states.

The distinction between the psychological and the phenomenal aspects of mental states is important because they indicate two distinct aspects to which agents may respond. For instance, if I desire a coffee, then I may desire it because I like the experience of tasting the coffee (a phenomenal property) or I may desire it because I want weaken my motivation to fall asleep (a psychological property). Or, to take a morally salient example, if I desire to relieve suffering, then the content of this desire may be the state of affairs in which agents do not *experience* suffering, or it may be the state of affairs in which agents *act in ways characteristic of* suffering, such as writhing in pain and complaining.

It strikes me that when moral agents are motivated by others' experiential states, they are motivated in virtue of the phenomenal aspects of these states rather than the psychological ones. I desire the prevention of your suffering because I don't want you to experience suffering, not because I don't want to hear you complain about it.

This may be true of *all* moral motivation, even if the states to which agents respond are not generally considered phenomenal states, such as beliefs and desires. For instance, a preference satisfaction consequentialist may be more troubled by the fact that unsatisfied preferences feel unpleasant than the fact that they prevent agents from acting in accordance with their desires.[46]

That said, it seems impossible to respect others' autonomy, for instance, without being motivated at least in part by the psychological aspects of their mental states. To respect one's autonomy is to respect their ability to choose for themselves how to act, an ability that is mediated by the psychological aspects of their beliefs and desires, not by how these beliefs and desires feel.

That aside, my point here is that motivation by moral reasons involves responding to the phenomenal aspects of mental states in a significant number of cases. This is relevant because unlike the psychological aspect of mental states, which plays a functional role in our behaviour and cognition, the phenomenal aspect is not obviously representational.[47] I say 'not *obviously* representational' because there are a wide range of theories about phenomenal mental states, some of which explain these states in terms of representational properties, but it's not at all obvious which of these theories is true.[48]

These are theories of consciousness. Although 'consciousness' is an ambiguous term,[49] one of the more philosophically interesting properties picked out by this term is the phenomenal aspect of mental states. The task of analysing these properties in terms of simpler non-phenomenal properties is often referred to as the hard problem of consciousness, as opposed to the easy problem of consciousness, which is the task of analysing psychological mental states, such as beliefs and desires. David Chalmers, who coined these phrases, notes that the easy problem is not easy, just easier than the hard

[46] Peter Singer (1979), for instance, is a preference satisfaction consequentialist who seems primarily concerned with experiential states, such as pain. For Singer, preference satisfaction consequentialism seems to be a more plausible alternative than hedonistic consequentialism as it allows us to make better sense of our intuition that the death of persons is worse than that of non-persons, given that non-persons are not self-aware, and therefore cannot have a preference for their own continued existence, unlike persons. Other consequentialists, including Jeff McMahan (2002), reject the intuition that the death of persons is worse than that of non-persons, on the basis that preference satisfaction consequentialism is less plausible than hedonistic consequentialism.
[47] See, for instance, Chalmers (1996)
[48] Schwitzgebel (2017)
[49] van Gulick (2021)

problem.[50] Explaining beliefs and desires in terms of simpler properties, such as representations, or in terms of physical processes, such as the activity of neurons, is easy because we can use empirical tools to test our predictions. By contrast, since phenomenal states are characterised by how they feel from the perspective of a single individual, and seem to be inaccessible to third-person observation, they seem almost impossible to study empirically and are thus far harder to explain.[51]

Given this, there are far fewer empirical constraints constraining the range of theories about phenomenal properties than there are constraining the range of theories about psychological properties, and thus there is a far greater range of theories about what phenomenal properties are.[52] As such, I am not comfortable assessing the plausibility of the main contenders, and I will mention only two such theories to give a taste of the range with which we are dealing.

At one end of the spectrum, some philosophers hold that phenomenal properties are a fundamental property of the universe, alongside fundamental physical properties such as mass and charge. These theories are generally forms of *panpsychism*, because they generally posit that, like mass and charge, phenomenal properties are found in almost all physical objects, from galaxies down to subatomic particles.[53]

At the other end, some philosophers hold that phenomenal properties only arise out of systems that can represent other mental states in the same system. These *higher order thought* theories suggest that phenomenal properties require not only representation but also metarepresentation, and thus restrict their presence to individuals capable of metarepresentation. On such theories, it would seem that neither human infants nor many nonhuman animals could have first-person experiences.[54]

Both of these theories, along with a range of theories in between, are plausible contenders for the correct theory of phenomenal properties.[55] Given this, it is not at all clear whether phenomenal properties are necessarily representational. Fortunately, this does not matter too much. Before I go on to explain why, I will briefly discuss the other type of mental state that can motivate agents to act morally, reactive attitudes.

Reactive attitudes include a range of emotional responses to the behaviour of other individuals and of oneself, including resentment, indignation, guilt, pride, and gratitude.[56] Because they are emotional responses, reactive attitudes exemplify the dual nature of many mental states, as possessing both psychological and phenomenal properties. Guilt, for instance, often motivates us to make amends but it also feels unpleasant in a way that is clearly distinct from other unpleasant feelings, such as pain or fear.

While the motivational push of emotions is a desire-like quality, many theorists also argue that they represent our bodily states, such as the sinking feeling in one's stomach

[50] Chalmers (1996)
[51] Chalmers (1996), Schwitzgebel (2017)
[52] Schwitzgebel (2017)
[53] Although I don't know whether he explicitly endorses panpsychism, Chalmers (1996) does offer an overview and defence of the theory.
[54] Peter Carruthers (2000) is perhaps the best known proponent of a higher order thought theory of consciousness. He has also explicitly speculated that nonhuman animals lack phenomenal mental states, although he has walked back this theory in recent years (2019)
[55] Schwitzgebel (2017)
[56] Strawson (1962)

characteristic of guilt.[57] In this respect, emotions are more like beliefs, aiming to represent the world as it is. Whether emotions do represent bodily states or whether they simply motivate individuals to act in particular ways, it seems clear to me that they generally do involve representation of some kind.

The phenomenal aspect of emotions, such as the unpleasantness of guilt or the experience of one's stomach sinking, strike me as distinct from the psychological aspect, which asks questions such as 'what is guilt for?' and 'does guilt represent bodily sensations?'. Explanation of these phenomenal aspects seems to be a hard problem, just like explanation of phenomenal properties in general.

So it makes sense to ask whether morally motivated agents who respond to others' reactive attitudes are responding to the psychological properties or the phenomenal properties. When I accept my friend's apology for standing me up, I am responding to her reactive attitude of guilt. But is my acceptance of her apology motivated by my desire that she no longer *feel* guilty or is it motivated by my desire that she *remains motivated* to continue our friendship? In my experience, it's probably a little of both.

So given that moral motivation seems to involve responses to both the psychological and the phenomenal aspects of others' mental states, and given that psychological aspects generally involve representation, whereas the jury is out on whether phenomenal aspects involve representation, does this make it possible that moral motivation *always* involves metarepresentation? I don't think so. The reason for this is that *even if* phenomenal states are necessarily representational, *representations of* phenomenal states need not represent these states *as* representational. Similarly, an infant can look at a picture of the beach and represent certain aspects of it – the colour of the paint on the page, the similarity of the beach to a real beach – without representing the picture *as* a representation of the beach. Thus, it is possible to represent mental states, and in particular to have desires about others' mental states, without metarepresenting, even if the mental states in question are representations.

This raises an important question. Which level of representation do agents use – primary, secondary, or metarepresentation – when representing the morally relevant aspects of others' mental states? Phrased differently, when agents respond to others' mental states in morally relevant ways, are they *necessarily* responding to representational aspects of those states, such that an agent without the ability to metarepresent would not be able to respond in the appropriate ways? When it comes to respect for autonomy or acting to ensure fairness, I think the answer is yes, agents must metarepresent to respond in the appropriate ways.

Failure to respect others' autonomy is bad because doing so interferes with the goals they have set for themselves. Given that these goals are a product of their beliefs and desires, one could fail to respect another's autonomy by interfering with their beliefs, as is the case when we lie to others, or by interfering with their desires, as is the case when we coerce others to do things that they don't want to do. Respect for autonomy involves sensitivity to these mental states and their role in allowing agents to pursue their goals. Given that this role requires the use of representations, respect for autonomy requires sensitivity to the representational aspect of beliefs and desires, and thus requires agents to exercise their ability to metarepresent.

[57] Antonio Damasio (1994) is perhaps the best known proponent of this view.

Sensitivity to fairness likewise involves sensitivity to the representational aspects of others' mental states. As noted in Chapter One, the reason for this is that sensitivity to fairness requires one to distinguish between things whose distributions *matter*, such that some distributions are fair while others are unfair, and things whose distributions *does not matter*, such that any distribution of these things will be neither fair nor unfair. Examples of the former are things like food and money, while examples of the latter are things like the total number of atoms in one's body. The ability to make this distinction relies on understanding the fact that things that matter are *desirable*, whereas things that don't matter are not. If one is motivated by fairness, as I am when I share things equally between my two children, then one must see these things as desirable, as the kinds of things that feature in the content of others' desires, and thus, one must be sensitive to the representational aspects of these desires.

Note that *moral judgement* is structurally analogous to motivation out of concern for fairness. Just as concern for fairness involves a desire about desirability, moral judgement involves a belief about desirability. In both cases, a higher-order representational state represents a situation as *desirable*. As discussed in Chapter Three, to represent something as desirable is to represent it as the *content of a desire*. Of course, the relevant desire need not actually exist; it is a hypothetical idealised desire, but its content is represented as content regardless. The implication here is that *flexible moral agency requires metarepresentation*, as too does moral motivation out of concern for either autonomy or fairness.

Concern for welfare may also involve sensitivity to the representational aspects of others' mental states, particularly if we take a pluralist view of welfare, as do objective list theorists. If knowledge is one thing constitutive of welfare, then ensuring that other agents have knowledge involves understanding the representational content of their beliefs. If relationships are one thing constitutive of welfare, then ensuring that other agents have relationships characterised by positive reactive attitudes is important, and this may involve sensitivity to the goal-directed aspects of such attitudes. Even some monadic accounts of welfare, such as preference-satisfaction accounts, may involve sensitivity to representational aspects of others' mental states, such as the goal-directed nature of preferences.

But what of experiential mental states? These are *wholly* constitutive of welfare on hedonistic accounts, and may also be partially constitutive of welfare on preference-satisfaction and objective list accounts. Does moral motivation by experiential states necessarily involve representation of their representational properties?

It strikes me that the answer must be no. Although higher order thought theories of consciousness conceive of experiential states as necessarily representational, the existence of other non-representational theories of consciousness shows that it is *possible* to conceive of experiential states as non-representational. This is perhaps most obviously true for panpsychism, according to some versions of which consciousness is a fundamental property of the universe, akin to gravity and electromagnetism, rather than a representation of something else.[58] If one believed that suffering were similarly non-representational, then one may be motivated to prevent another person's suffering without being motivated by any representational aspect of this suffering.

More prosaically, when I give my children medicine to relieve their suffering, I don't do so because their experience of suffering *represents* some bodily disequilibrium (although it may) and because medicine would remove the source of the disequilibrium. Rather, I give

[58] For example, Chalmers (1996).

them medicine because I want their suffering to *go away* and I believe that if I give them medicine then their pain will go away.

One may wish to claim at this point that even if the desire does not involve metarepresentation, metarepresentation is still involved because of the structure of the *belief* involved. Specifically, conditional beliefs, those of the form "I believe that if X then Y", involve representation of the relationship between X (I give my children medicine) and Y (my children's pain will go away). The representation of this relationship may lead one to think that conditional beliefs involve metarepresentation.

But although metarepresentation does involve the representation of a relationship, not all representations of relationships are cases of metarepresentation. If I believe that the person standing in front of me is my mother, I do not thereby *metarepresent* her as my mother. Rather, I *represent* her as my mother. To count as metarepresentation, the relevant relationship must be one of *representation* and not one of, for instance, motherhood. By contrast, if I believe that the photograph in front of me represents my mother, then this is a genuine case of metarepresentation. Consider another conditional belief: a toddler may believe that if she looked inside a box of chocolates then there would chocolates inside. If she were to find crayons instead then she would be surprised, which would suggest the presence of this conditional belief. Indeed, this is what psychologists have found.[59] But, again, she merely *represents* the box as containing chocolates. She does not metarepresent because she does not represent the box as representing the chocolates. The relevant relation between the box and the chocolates is one of containing, not one of representation.

Finally, one may believe that responding to concerns of suffering necessarily involves metarepresentation because it involves representing another's suffering as *bad for them*.[60] If this is the case, then perhaps my explanation of my giving my children medicine to relieve their suffering is mistaken. Instead of involving a belief that giving them medicine will alleviate their suffering and an *intrinsic* desire to alleviate their suffering, without representing this suffering as bad for my children, perhaps this desire is in fact instrumental, derived from a belief that suffering is bad for my children and a desire to shield my children from badness in general.

This strikes me as overthinking things. As a general point, although agents typically desire good things and desire to avoid bad things, one need not understand the concepts of goodness and badness to be so motivated. Even those animals with relatively simple nervous systems, such as worms and insects, move toward food and away from predators. Assuming that these movements are mediated by representations of their environment, it nonetheless seems implausible to claim that these creatures represent the goodness of food and the badness of predators. Rather, it seems that they represent their presence of food and predators and act accordingly.

One may still claim that to be motivated by suffering involves representation of this suffering as bad for the victim, because suffering is *valanced*. When I look at the sky, for instance, I have a visual experience of the colour blue, but I don't usually experience this as positive or negative. If I were to look at the sun, however, I would not only have a bright yellow visual experience, but also a strongly negatively-valanced experience of discomfort. I experience the sky as blue but I experience the sun as yellow and *as bad*. It might be thought that to experience something as bad is to experience it as *bad for oneself*. But again, there's no obvious reason to hold that experiences of badness are necessarily

[59] Perner (1991)

[60] Thanks to my supervisor Garrett Cullity for this suggestion.

accompanied by experiences of badness for oneself, especially in the case of neurologically simple animals that seem to feel pain but seem to lack any sophisticated kind of self-consciousness.

Similarly, it is not obvious that motivation by *others'* suffering is any different. It strikes me that one can represent another's suffering, without necessarily representing it as being bad for the victim. Having said this, however, this seems no more obvious than the alternative. Perhaps representing another's suffering necessarily involves representing it as bad, and perhaps this necessarily involves representing it as bad for the victim.

Even so, there is good reason to think that this does not require metarepresentation. As is the case with conditional beliefs, representation of suffering as bad for the victim involves representation of *a* relationship but it does not involve representation of the *right kind* of relationship to count as metarepresentation. If I believe that suffering is bad for my children, I do not thereby *metarepresent* suffering, badness, my children, or the relationship between them. Rather, I merely *represent* all these things. The relevant relationships are not relationships of representation but of prepositional relations (my children are *in* pain), the possession of properties (pain *is* bad), or combinations of the two (pain *is* bad *for* my children).

Thus, concern for welfare involves responding to *non*-representational aspects of mental states, particularly if one accepts a hedonistic account of welfare but even if one accepts a preference satisfaction or objective list account in which experiential states are partially constitutive of welfare. It follows, then, that if moral motivation involves concern for welfare, either as the only foundation of morality or as one foundation in a pluralist account, then it does not require metarepresentation.

Moral motivation only *necessarily* involves metarepresentation if welfare is not a foundation of morality. I find such a view implausible but perhaps some Kantians or contractualists may disagree. On such views, basic moral agency would thereby require metarepresentation.[61] If one accepted such a view then one would be committed to moral agency occurring later in childhood development than if one accepted welfare as a foundation of morality.

Given that concern for welfare does not require metarepresentation, one can then ask whether it involves secondary representation or whether primary representation is sufficient. To be clear, concern for welfare must involve *some* level of representation, because it requires a desire *about* the relevant experiential state, which is to say that it requires *representational content* of the relevant experiential state.

It strikes me that I can be concerned for *my own* welfare without the ability to use secondary representation. Recall that the distinction between primary and secondary representation is one of content availability: secondary representation is characterised by representational models that can draw on content from other representational models, whereas primary representation is characterised by models that cannot.

Now suppose I want a coffee because I enjoy the taste. The content of my desire is the state of affairs in which I have the experience of the taste of coffee. Of course, one would not specify the content of the desire this specifically in everyday contexts, but if the coffee

[61] And flexbile moral agency would then seem to require second-order metarepresentation, or *meta-metarepresentation*. Empirical evidence from second-order false belief experiments suggests that this develops between the ages of five and six years (Perner 1991).

were burnt and I missed out on experiencing the desired taste then it seems clear that my desire would be unfulfilled.

To be motivated by a concern for my own welfare in this case is simply to act upon my desire for the experience of having a good-tasting coffee. The content of this desire need not have any connection to any other representational model. I simply have the desire and I act upon it. Having done so, I have the positive experience, which constitutes an improvement in my welfare. No secondary representation required.

Of course, being motivated by concern for *one's own* welfare is not moral motivation. Being motivated by concern for *others'* welfare is, so it is important to investigate whether this requires secondary representation or whether primary representation is sufficient. Here I consider two phenomena: *object permanence* and *emotional mirroring*. Both phenomena seem related to the representation of others' experiential states: object permanence seems to share some important structural similarities, while emotional mirroring seems to directly involve the representation of others' experiential states. Moreover, both develop in early infancy, suggesting that neither involves secondary representation.[62] Nonetheless, I shall argue that neither adequately explain the ability to represent others' experiential states, and that the best explanation for this ability involves secondary representation. The two arguments are independent, but taken together they lend strong support to the claim that secondary representation is necessary for moral motivation.

Object permanence typically develops over a period from four to eight months of age and is characterised by the expectation that objects continue to exist when unseen.[63] Developmental psychologists use an expectation violation paradigm when testing for psychological abilities, such as object permanence, in preverbal infants. This paradigm relies on the fact that 'surprising' phenomena will hold our (and infants') interest for longer, so we will look for longer at a phenomenon when our expectations are violated. In the case of object permanence, this is tested by showing an object to infants, occluding the object so that it is no longer in view, either removing the object or leaving it in place, and then removing the occlusion. Infants under four months of age show no difference in how long they look in either case, whereas infants over eight months consistently look for longer in the case where the object was removed. This is taken as evidence that these older infants expect the object to persist and are surprised when it seems not to.

Object permanence is a development *within* primary representation because it does not involve representations of two distinct scenarios, just a single scenario in which not every object is represented as part of one's immediate environment. Nonetheless, it seems to be an important precursor to secondary representation because it allows for the representation of the objects in two distinct *ways*: as present and as absent.

Others' experiential states are in some respects like occluded objects. I cannot see whether another person experiences pain. I know that such experiences exist, because I have felt pain myself, but others' experiences of pain are inaccessible to me. To flesh out the analogy, suppose that a nine-month-old infant is holding a blueberry, which she gives to me and which I hide in my hand. She previously believed that the blueberry was in her hand and now she believes that the blueberry is in my hand. Now suppose that this infant has dropped an object on her hand, causing her pain. She later sees me drop the same

[62] Bornstein (2013)
[63] *Ibid.*

object on my hand. Just as she can't see the blueberry, yet she believes that it is in my hand, does she now believe that there is pain in my hand?

I'm inclined to think that the analogy breaks down here. Unlike the blueberry, the pain in my hand is not the *same* pain as in the infant's hand. In the case of the blueberry, the child watches as I hide it and she represents the blueberry in my hand as the same blueberry. It's not clear that she can do this for pain, because she cannot see pain move from herself to others. She simply experiences pain come and go *in her own hand*. When I experience pain in my hand, she cannot represent the pain as moving from her to me in the same way as she does with the blueberry. If her hand still hurts, then there is no transfer. The pain is still in *her* hand, just as if the blueberry were still in her hand. And if her hand no longer hurts, then the pain has gone away and 'reappeared' in my hand, without being represented as having done so.

It may be that my pain isn't represented as the same pain, but as other pain. Just as there may be other, unseen blueberries in the fridge, there are other unseen pains in other people. But can an infant *represent* these unseen pains? If this is the case, I don't think object permanence provides any evidence for this. When an infant represents an object as persisting despite being visually inaccessible, there is no evidence that she represents it *as a distinct object* from its former, visually accessible self. It's not as if the infant sees a blueberry, watches as it is occluded by a screen, and goes on to represent a different blueberry in its place. Object permanence does not grant infants the ability to 'perceive' new objects with which they have no prior familiarity, even if such objects are qualitatively similar to familiar objects. For these reasons, the ability to attribute experiential states to others seems to require secondary representation.

Now consider emotional mirroring. It is common knowledge that preverbal infants can copy others' emotional expressions.[64] If I smile, then my infant daughter will often smile back. Of course, since infants cannot tell us how they feel, we can only infer their experience from their emotional expressions. It may be the case that mirrored expressions are not accompanied by the relevant emotional state, but my intuition is that this is not the case. In my own experience, when I see others' emotional expressions, I often feel a little of the emotion and I often mirror the corresponding expression. I see no reason why this would not generalise to infants.

But if my daughter feels joy after smiling back at me, does she represent *my* joy? To do so would involve not just representation of the experience of joy, but representation of this experience *as mine*. This seems to misdescribe what is going on here. Rather, it seems that she has a visual representation of my smile and an experience of joy. It is not obvious whether this experience *is* a representation (as representational theories of consciousness would claim), whether it is the *content* of a representation (as is the case with my desire to enjoy a cup of coffee), or whether it is wholly uninvolved in any representation. As previously discussed, the former two cases require only primary representation, whereas the latter case does not require representation at all. If secondary representation is involved at all, it would seem to operate in the causal connection between the visual representation of my smile and the experience of joy.

It is possible to explain this connection using secondary representation. Recall that perspective-taking involves integrating information from a secondary representational model into the primary model. Specifically, the secondary model is a model of the world

[64] Bornstein (2013)

from another perspective, and the primary model is the individual's model of the way the world actually is. I have discussed this in the context of mirror self-recognition, but a different example is more relevant here: *deception by simulation*. It is a well-documented fact that chimpanzees engage in deception. For instance, lower-ranking chimpanzees will often try to hide food from higher-ranking members of the group by putting it in places that are visible to themselves but not to their competitors.[65] It is also well-established that chimpanzees are capable of secondary representation but not metarepresentation: they pass the mirror self-recognition test but not the false belief test.[66] Chimpanzees' ability to deceive their competitors is explainable in much the same way as their ability to pass the mirror self-recognition test. They integrate information from a secondary model – in this case, the model of the world from their competitor's perspective – into their primary model of the way the world actually is.

Something similar seems to occur in our emotional lives. In particular, it strikes me that we often infer others' emotional states by simulating them, by representing the world from their perspective. For instance, when I see others' emotional expressions, such as the wide-eyed fear expression, for instance, I often feel a little of the emotion myself and I come to believe that the other person is feeling that way. We might call this ability *empathy by simulation*.

But given that this explanation invokes secondary representation, it seems to be misplaced when applied to infant behaviour. So how else might we explain the relationship between my smiling and my daughter feeling joy? A plausible explanation is that infants come prewired with certain desires: for food, to be free of pain, for attachment with others, and so on, and that these desires have a distinct phenomenal character both when present and when fulfilled.[67] For instance, when the desire for attachment is fulfilled – say, by being smiled at, – it may be *experienced as joy*. But while this explanation does not invoke secondary representation, nor does it invoke representation of another's experiential states. Thus, emotional mirroring, as with object permanence, does not provide evidence that representation of others' mental states can occur without secondary representation.

Since moral motivation involves secondary representation, we should briefly discuss this. Specifically, it implies that the representational content of the relevant desire – others' mental states – is made available to that desire from another mental model.

We have already encountered one case in which a desire seems to derive content from another representational model: the toddler's desire to do things for oneself. I suggested that such desires can exist because secondary representation makes it possible for representational content to be shared between models; in this case, representational content from a model describing what one *is* doing is made available to a model describing what one *wants* to do. In this case, it is possible for the toddler to desire to do things for herself because secondary representation makes it possible to have new desires whose content is derived from other mental models. Since she can have beliefs about doing things for herself, secondary representation makes it possible for her to have desires with similar content.

It is likely that empathy by simulation enables moral motivation in much the same way. Empathy by simulation provides a mental model about what others are feeling, and secondary representation enables content to be available to other models, including desires

[65] Byrne & Whiten (1988)
[66] Suddendorf & Whiten (2001)
[67] Schroeder (2004)

about others' feelings. A toddler, who is able to mentally represent her parents' happiness, may come to desire that her parents be happy.[68] As argued in Chapter One, agents are praiseworthy for acting on such desires, all else being equal.

More empirical research is needed to determine whether toddlers are capable of moral motivation, but the conceptual argument given above suggests that they are. In short, secondary representation makes it possible for representational content to be shared between discrete mental models, as indicated by mirror self-recognition, pretend play, emotional meltdowns, and wanting to do things for oneself, all of which typically develop between the ages of one and two years.[69]

In the final section, I will consider in more detail the evidence indicating when secondary and metarepresentation develop in children, as well as the prevalence of secondary and metarepresentation in nonhuman animals. This will then give a lower bound for the ages at which children become moral agents, and for the possibility of its existence in animals.

Evidence from Developmental and Comparative Psychology

I have claimed that secondary representation develops between the ages of one and two years, and that metarepresentation develops between three and four years. In this section, I will examine the empirical evidence for these claims. In particular, I will discuss research into the emergence of four specific abilities in childhood development: the *attribution of false beliefs* and *inhibition of behaviour*, both of which are associated with metarepresentation, and *pretend play* and *mirror self-recognition*, both of which are associated with secondary representation.[70] As we shall see, the developmental evidence bears this out. There has also been research into whether nonhuman animals have these skills, especially false belief attribution and mirror self-recognition. This research has failed to show that animals are capable of the former but that a few species are capable of the latter. Taken together, the empirical evidence suggests that only these species, as well as humans over the age of one, are capable of secondary representation, while metarepresentation is restricted to humans over the age of three. If I am correct that moral motivation requires secondary representation and moral judgement requires metarepresentation, then this provides an outer bound on which agents could be basic moral agents, and which could be flexible moral agents.

Let's consider false belief attribution. The false belief test was initially proposed by philosopher Daniel Dennett[71] in a commentary on a 1978 article by primatologists David Premack and Guy Woodruff,[72] which posed the question of whether chimpanzees understood the minds of others. Dennett suggested that scientists could test for this by checking whether subjects could attribute false beliefs to others, on the basis that attributions of *true* beliefs may be indistinguishable from one's own understanding of the situation and attributions of *desires* may be indistinguishable from understanding others' behaviour, rather than their mental states. Hans Wimmer and Josef Perner used a version of this experiment to test whether young children understand false belief, and published their results in 1983.[73] In their experiment, they told a story acted out with dolls and props.

[68] Schroeder (2004)
[69] Perner (1991), Doherty (2007)
[70] Perner (1991), Doherty (2007)
[71] Dennett (1978)
[72] Premack & Woodruff (1978)
[73] Wimmer & Perner (1983)

As the original was published in German, Martin Doherty offers this translation of the original story:

"Maxi is helping his mother to unpack the shopping bag. He puts the chocolate into the green cupboard. Maxi remembers exactly where he put the chocolate so that he can come back later and get some. Then he leaves for the playground. In his absence, his mother needs some chocolate. She takes the chocolate out of the green cupboard and uses some of it for her cake. Then she puts it back, not into the green cupboard but into the blue cupboard. She then leaves to get some eggs and Maxi returns from the playground, hungry.

Test question: Where will Maxi look for the chocolate?" [74]

Wimmer and Perner found that children began to pass this test at around four to five years of age, whereas younger children tended to answer that Maxi would look in the blue cupboard, where the children themselves believed the chocolate to be. Subsequent research into theory of mind has largely focused on this and other false belief tests, and has consistently shown that children under the age of three and a half years do not pass the test at rates above chance.[75]

In 1999, Joseph Call and Michael Tomasello adapted the false belief test for use with chimpanzees.[76] In their version of the experiment, they first ran a familiarisation task in which an experimenter hid food in one of two boxes, out of the chimpanzee's sight, after which a second experimenter placed a wooden block on top of the box containing the food, to indicate to the chimps where the food was.

Once the chimpanzees understood the familiarisation task, the experimental setup was altered to test for chimpanzees' understanding of false belief. In this task, the second experimenter, in view of the chimpanzee, *watched* as the first experimenter placed food into the box, which was still out the chimpanzee's sight. Then the second experimenter left and the chimpanzee watched as the first experimenter switched the location of the boxes. When the experimenter returned, he placed the block on the empty box, where he should think that the food was.

If the chimpanzees could attribute false beliefs to the first experimenter, then they should have chosen the box without the block on it. Instead, they consistently chose the box with the block.

This suggests that chimpanzees cannot attribute false beliefs, and thus fails to provide evidence that chimpanzees are capable of metarepresentation. For this reason, I tentatively conclude that chimpanzees are not capable of moral judgement. In fact, I am not aware of any research indicating that any nonhuman animal passes the false belief test or any other evidence that they are capable of metarepresentation,[77] so it seems plausible to conclude that flexible moral agency is limited to human beings above the age of three and a half years.

---

[74] Doherty (2007)

[75] For instance: Baron-Cohen *et al.* (1985), Hogrefe *et al.* (1986), Perner *et al.* (1987), Gopnik & Astington (1988), Moses & Flavell (1990). Wellman *et al.* (2001) also performed a meta-analysis of 77 reports or papers from 1983-1998 and provided an estimate of the probability of children passing the false belief test at various ages. They concluded that children were 50% correct at 3 years 8 months. However, Onishi & Baillargeon (2005) have since published research indicating that children as young as 15 months can attribute false beliefs. I will discuss this research below.

[76] Call & Tomasello (1999)

[77] Whiten & Suddendorf (2001)

There is however, some reason to believe that children's ability to metarepresent emerges much earlier than this. In an experiment published by Onishi and Baillargeon[78] in 2005, experimenters tested the ability of younger children to pass the false belief test, using an experimental protocol designed for use with nonverbal infants, and found that children as young as fifteen months passed this version of the false belief test.

This *violation of expectation* (VOE) protocol relies on the fact that unexpected phenomena are surprising, and thereby hold our attention for longer than expected phenomena. Thus, if infants can attribute false beliefs to others, we should expect them to be surprised, and thereby look for longer, if they see an agent look for a hidden object in a location that does not correspond to their (false) belief.

The experimental setup was similar to that of Call and Tomasello. In the familiarisation phase, an experimenter hid food in one of two boxes, which was moved to the other box once the experimenter left the room. In the experimental phase, the experimenter returned and looked in one of the two boxes. The authors predicted that if children could attribute false beliefs to others then they would be surprised, and therefore look for relatively longer, when the experimenter looked in the box that did not correspond to their (the experimenter's) false belief. Onishi and Baillargeon found that children as young as 15 months looked for longer in such cases, and concluded from this that they have an understanding of others' false beliefs.

Subsequent experiments using a different *anticipatory looking* (AL) protocol have corroborated this finding. The AL protocol, like the VOE protocol, relies on interpreting infants' eye movements as indicating their expectations. Where it differs is that it involves recording where infants look *before* an event occurs, which is taken to show that the infant expects something to happen in this location. In the case of false belief experiments using this protocol, researchers have found that infants tend to look in the location corresponding to the agent's false belief.

In a 2018 commentary on a special issue responding to the literature on false belief attribution in infants, Baillargeon and her co-authors reported that the results from VOE and AL experiments were well-replicated, with over 30 published studies providing independent evidence for the findings of the initial 2005 study.[79] Thus, we cannot ignore this body of literature.

However, in a 2019 meta-analysis, Barone, Corradi, and Gomila found that of the studies reporting false belief understanding in infants, the earlier studies showed a greater effect but had small sample sizes, whereas later studies with larger sample sizes showed a much more modest effect, suggesting that these findings have been difficult to replicate. They also used statistical models to argue that there may be publication bias with regard to these findings.[80]

That aside, if the findings from VOE and AL tasks are accurate, they present a paradox. How is it that 15 month old infants pass these *implicit* false belief tasks but older children fail traditional *explicit* false belief tasks until shortly before their fourth birthday? One suggestion in defence of the claim that infants do understand false belief is that traditional false belief tasks are more taxing on children's executive function skills and are therefore more difficult to pass. This is a plausible explanation but is undermined by studies showing

[78] Onishi & Baillargeon (2005)
[79] Baillargeon, Buttlemann, & Southgate (2018)
[80] Barone, Corradi, & Gomila (2019)

that executive function is correlated with performance on theory of mind tasks, regardless of whether these tasks impose high or low executive demands.[81]

A more plausible explanation, I think, is that even if these findings are accurate, they do not yet show that infants are capable of attributing false beliefs. Or, rather the experiments do not show that infants attribute to others the *representational properties* of false beliefs. Many critics of these experiments concur. They claim that while infants demonstrate expectations in VOE and AL experiments, the content of this expectation is not about the representational properties of agents' false beliefs, but rather simpler facts that need not involve metarepresentation.

For example, Cecelia Heyes has claimed in a 2014 critical review of 20 implicit false belief experiments, that the results of these experiments could be equally well-explained by infants forming expectations about features of the experimental setup, rather than expectations about agents' false beliefs. Heyes gives alternative interpretations of each of the studies in her review, covering seven different experimental setups, including both VOE and AL protocols. In all cases, Heyes claimed that infants responded to perceptual or imaginative novelty introduced in the experimental phase of the relevant experiment, which distinguished this phase of the experiment from the earlier familiarisation phase.[82]

Another low-level explanation is that infants track certain rules about behaviour rather than false beliefs.[83] One such rule, which is salient in many false belief tasks is that agents tend to look for objects in the last place they left them. One can understand this rule without understanding the mental states that cause this behaviour.[84]

Ted Ruffman notes two skills possessed by infants at or near birth, which make it possible for them to predict agents' behaviour directly, without the attribution of mental states. First, he notes that infants have an innate capacity for statistical learning that makes it possible for them to draw inferences from repeated exposure to phenomena. Second, he notes that infants are especially attuned to the human faces and the movements of humans and can discriminate between human and nonhuman movements from four days of age.[85] Given that these skills are present at or near birth, it is plausible that by 15 months these skills develop to the point where it is possible for infants to predict agents' behaviour in a way consistent with understanding false belief but without actually having this understanding.

For these reasons, I remain sceptical that children understand the representational aspects of false belief, or are capable of metarepresentation generally, before about three and a half years of age. This is supported by another line of evidence from experiments on children's executive function. One such task presented children with a variation of the 'Simon says' game, in which two toy animals gave instructions to children. The children had to follow the instructions given by the toy elephant but not those given by the toy bear. Children from the ages of 3 years 0 months to 3 years 2 months performed poorly, following the instructions of both characters, whereas children from 3 years 3 months to 3 years 5 months passed 76% of trials.[86]

---

[81] For instance, Carlson, Claxton, & Moses (2015)
[82] Heyes (2014)
[83] For instance, Perner & Ruffman (2005), Apperly & Butterfill (2009), Ruffman (2014), Perner (2014)
[84] Perner and Ruffman (2005)
[85] Ruffman (2014)
[86] Jones, Rothbart, and Posner (2003), discussed in Doherty (2007), p. 131

It strikes me that this ability to inhibit actions requires metarepresentation because it involves two contradictory desires, the desire to win the game and the desire to follow the bear's instructions, the belief that the content of the first desire is preferable to that of the second (that is, the belief that winning the game is preferable to following the bear's instructions), and crucially, the belief that one ought to suppress one's desire to follow the bear's instructions in order to win the game.

One could imagine a similar scenario playing out in which the bear and elephant were replaced by one's sibling and oneself. My sibling asks me to share my toy but I don't want to. At the same time, I want my sibling to be happy, so I have competing desires. If I am able to inhibit my own desire to keep the toy for myself then I can share the toy. Thus, the ability to inhibit action, which seems to require metarepresentation, seems to enable moral judgements. The fact that this ability develops at around the same time that children pass the false belief test gives further support to the claim that they are both manifestations of the same ability to metarepresent, and that this ability emerges between the ages of three and four years.[87] Given that moral judgement requires metarepresentation, this implies that the lower bound for flexible moral agency is no earlier than three years of age.

Turning now to basic moral agency, we can ask the same questions. Given that basic moral agency requires secondary representation but not metarepresentation, it is plausible that it develops earlier than flexible moral agency. The empirical evidence indicates that secondary representation emerges at between one and two years, which puts a lower bound on the development of basic moral agency.

Given that secondary representation is simpler than metarepresentation, it would not be too surprising if it were present in some nonhuman animals. If so, then this suggests the possibilities that these animals could be basic moral agents. Although this may seem farfetched, this is not obviously mistaken. Mark Rowlands, for instance, has written on the ability of animals to respond to moral reasons and while he does not conclude that these animals are moral agents, he does claim that some animals can respond to moral reasons and thus are deserving of a certain kind of admiration in virtue of this. It strikes me that Rowlands's conclusion that animals are not moral agents rests primarily on the fact that he adopts a more restrictive definition of moral agency than I offer here.[88]

The clearest evidence given for the presence of secondary representation in the theory of mind literature is that of pretend play in toddlers. Pretend play is characterised by children acting as if certain nonreal situations were real, such as pretending that a banana is a phone, pretending that today is my birthday, or using a nonexistent spoon to eat nonexistent food.[89] Unlike false belief tests or tests for executive function or mirror self-recognition, no such test is needed for the presence of pretend play. It is an almost ubiquitous feature of early childhood, and is well-known to develop between about 12 and 18 months.[90] As

[87] Although the empirical evidence that metarepresentation emerges after the age of three is remarkably consistent (with the controversial exception of implicit false belief tasks), it is also the case that communication difficulties can delay its development. In particular, autistic children and deaf children of hearing parents can face significant delays, as found by Peterson & Siegal (1999), Schick *et al.* (2007), Russell *et al.* (1991), Baron-Cohen *et al.* (1985), and many others. Doherty (2007), pp. 155-177, 186-198, offers an accessible discussion of the research in these areas.
[88] Rowlands (2015)
[89] These examples refer to cases of *object substitution*, *attribution of pretend properties*, and *invention of nonpresent objects*, respectively (Doherty 2007, p. 92). However, this distinction need not concern us here, as each involves secondary representation in similar ways.
[90] Doherty (2007), p. 92.

such, the relevant question with respect to pretend play is not when it develops, but whether it is evidence of secondary representation.

Perner claims that pretend play does requires secondary representation because when the child acts as if certain nonreal situations were real, she represents a hypothetical situation, which she can easily distinguish from reality. He contrasts this with misrepresentation, which involves *unintentional* representation of nonreal situations as if they were real. Pretend play, on the other hand, involves the intentional creation of an imaginary situation and requires secondary representation to be able to distinguish this situation from reality. This knowledge is analogous to the knowledge one has of one's desires or of other perspectives, as is the case with emotional meltdowns and mirror self-recognition, respectively.[91]

From the other side, while pretend play can involve metarepresentation (particularly in slightly older children), Perner argues against it requiring metarepresentation. In short, pretend play involves a representational model of a pretend situation *as a pretend situation*, not as a representation of a pretend situation.[92]

Perner illustrates this distinction using sandbox models of a battlefield. In this example, generals use a sandbox – a model of the battlefield – to represent the known location of their own and enemy troops. If they wish to plan an attack, they use a second sandbox to do this. The reason for the second sandbox is so they don't confuse the actual locations of the troops with where they plan to send them. These sandboxes correspond to the real (primary) and pretend (secondary) situations, respectively. The miniature troops within both sandboxes represent human troops: those in the first sandbox represent real troops out in the field, while the ones in the second sandbox represent troops in the hypothetical attack scenario. Depending on the detail of the models, miniatures of specific identifiable soldiers may exist in both sandboxes. This correspondence does not imply, however, that the miniature troops in the second sandbox represent their counterparts in the first sandbox. As stated above the miniature troops in the second sandbox represent something else, namely the human troops in the hypothetical scenario.

Of course, however, the *generals* metarepresent by representing both sandboxes as representing human troops. But the task of distinguishing the two sandboxes and designating one as real and the other as hypothetical could be accomplished without *representing* one as real and the other as hypothetical. In general, metarepresentation *cannot* be required for the task of distinguishing different representations on the basis of their distinct functions because metarepresentation itself requires the ability to distinguish between representational content, such as the miniatures in the sandboxes, and the targets of this content, such as the human troops, and to understand that the former is meant to represent the latter. In order to do this, one must represent both the content and the target and to designate the former hypothetical and the latter as real.

Likewise, when a toddler pretends that a banana is a phone, she must represent both. But she need not represent the banana as a phone. Rather, she has a cognitive mechanism that designates the representation of the banana as real and the representation of the phone as pretend. Because these representations are designated in this way, she can act on the basis of these representations of pretend scenarios as if they were representations of real scenarios.

While pretend play is typically thought of as involving a hypothetical belief, such as the belief that the cloth is a pretend pillow, it strikes me as also involving desires of a certain

[91] Perner (1991), p. 51
[92] *Ibid.*, p. 53

kind. When a child pretends to answer the pretend phone, she is acting on a pretend desire. By 'pretend desire' I do not mean that no desire exists. Rather, I mean that she acts on a real desire whose content is specified by the pretend scenario. When a toddler pretends to sleep, for instance, she does not want to sleep but she wants to 'pretend' sleep.

The banana phone example should illustrate the relevance of 'pretend desires' to moral agency. When a child pretends to answer an imaginary phone, she acts on a desire to respond in a specific way ('talking') to something that she cannot see (a nonexistent conversational partner). As I mentioned in the previous section, responding to others emotions and experiential states likewise involves acting on a desire to respond in a specific way to something that one cannot see.

There has been less research on pretend play in nonhuman animals, but there is another line of research that suggests that at least some nonhuman animals are capable of secondary representation. In humans, mirror self-recognition develops at around the same age as pretend play and both involve secondary representation. This suggests the possibility that at least some nonhuman animals may be basic moral agents.

The classic test for mirror self-recognition was developed by Gordon Gallup and was used by him to determine that chimpanzees could recognise themselves in a mirror but that two species of monkey (rhesus macaques and stump tail macaques) could not.[93] The experiment was performed as follows. The experimental setup involved a 10-14 day familiarisation phase, in which the animals were exposed to a mirror for 8 hours each day. They initially exhibited social behaviours, treating their reflection as if it were another of their species. This remained the case for the monkeys, but the chimpanzees exhibited an increase in self-directed behaviour by the third day. This included picking food from their teeth and grooming otherwise visually inaccessible parts of their body while looking in the mirror. After the familiarisation phase, the animals were anaesthetised and marked with a red pigment on an eyebrow ridge and the opposite ear. After recovery, the chimpanzees, but not the monkeys, exhibited behaviour directed at the marks, such as touching them and inspecting their fingers afterwards. As an additional control, another group of chimpanzees were marked and tested without undergoing the familiarisation phase. These chimpanzees showed no special interest in the marks.[94]

Subsequent research on human children has indicated that they typically pass the mark test during their second year.[95] A longitudinal investigation of children at three-month intervals between 12 and 24 months found that children tended to pass the mark test at around 18 to 21 months.[96]

The test has also been conducted on other animals and has consistently found that both chimpanzees and orangutans pass the mark test, whereas no species of monkey do so.[97] In a 2017 review, Diana Reiss and Rachel Morrison report that mirror self-recognition has been well-documented in all species of great ape (chimpanzees, gorillas, orangutans, and

[93] Gallup (1970)
[94] As described in Tomasello & Call (1997), p. 331
[95] For instance, Amsterdam (1972), Anderson (1984)
[96] Nielsen *et al.* (2003)
[97] Tomasello & Call present the results of 53 studies from 1970 to 1996, which show that chimpanzees and orangutans have passed the mark test in multiple studies. While some studies show that some species of monkey also pass the mark test, these studies have failed to replicate (1997, pp. 332-333).

bonobos), but not in monkeys or gibbons. They also report that two non-primate species have passed modified versions of the mark test: bottlenose dolphins and Asian elephants.[98]

Many species have been tested on their reactions to mirrors. As far as I can tell, these studies have either shown that subjects fail the mark test, failed to administer the mark test, failed to replicate, or been criticised for methodological flaws.[99] That said, some species that fail the mark test have been observed to respond to mirrors in interesting ways. In particular, studies involving monkeys, dogs, pigs, corvids, and parrots have shown that these animals use mirrors to guide their behaviour, such as by using the mirrors to find hidden objects.[100]

I contend that passing the mark test is evidence of secondary representation, whereas other forms of mirror-guided behaviour are not. If this contention is correct, then we currently have evidence for secondary representation, which is required for basic moral agency, only in humans, other great apes, dolphins, and elephants.

I suspect, like the infants who pass implicit false belief tests, the animals that engage mirror-guided behaviour are capable of exploiting statistical regularities in the environment. In this case, they exploit the correspondence between the mirror reflection and the real world in order to use the mirrors to find hidden objects.

Importantly, the ability to use statistical regularities in this way need not require secondary representation. Many animals, including perhaps all vertebrates, are capable of associative learning, which exploits statistical regularities between various stimuli. For instance, the paradigmatic association of a ringing bell with the smell of food, manifested by the animal salivating upon hearing the bell,[101] involves integration of two *primary* representations: the olfactory representation of the food and the auditory representation of the bell.

Of course, animals vary in the flexibility of their associative learning. Some animals are capable of making associations only in limited circumstances, while others are capable of making novel associations between disparate phenomena.[102] Given this, and given the close correspondence between an animal's environment and the reflection of this environment in a mirror, it is unsurprising that some animals can make associations between their between the two. Nor is it surprising that some animals can draw inferences about hidden items on the basis of the general correspondence between their environment and its reflection, as animals are capable of finding hidden items on the basis of other, less obvious correspondences, such as the indirect and contingent correspondence between the presence of a male lyrebird and the sounds that comprise a lyrebird song, which vary widely and mimic other sounds in the environment. And yet, female lyrebirds are capable of learning these associations and using them to find male lyrebirds.[103] Of course, we should expect that species whose survival depends on their ability to draw associations should be able to do so. But the ability to make associations, even complex associations, can be achieved without secondary representation. These associations seem to be between multiple representations within a *single representational model*. The lyrebird song and the inferred male lyrebird belong to the same model of reality. The same seems to be true of the use of mirrors to find hidden items. In this case, the model represents both the location

[98] Reiss and Morrison (2017)
[99] *Ibid.*
[100] *Ibid.*
[101] Pavlov & Anrep (1928)
[102] Pearce (2013)
[103] Kaplan (2019)

of the hidden item and the item in that location, as inferred from the learned regularities between the environment and its reflection in the mirror.

Passing the mark test is different. Animals who pass the mark test have already habituated to the presence of mirrors during the familiarisation phase of the experiment, and have learnt to draw associations in much the same way as other animals who can use mirrors to find hidden items. Where apes, dolphins, and elephants differ, however, is in using mirrors to explore locations that cannot be visually accessed *at all* without the mirror. Gallup's chimpanzees had never seen inside their own mouths before, so there was no statistical regularity for them to exploit. Instead, it seems that chimpanzees (and other species that pass the mark test) have a primary model of the world from their own perspective and a secondary model of the world from the perspective of the mirror, and they use the associations between these *distinct representational models* to explore their own reflections.

Taken together, this evidence suggests that children from the age of about 12-18 months, as well as great apes, dolphins, and elephants, are capable of secondary representation.[104] Whether this implies that they are capable of moral motivation, and therefore basic moral agency, depends on whether they can specifically represent others' mental states. As I have argued, though, representation of others' mental states requires secondary representation, and many abilities associated with secondary representation develop in tandem. Moreover, given that representation of mental states offers a powerful way of explaining others' behaviour,[105] it is plausible that the highly social species capable of secondary representation would be able to represent and be motivated by others' mental states. Thus, I tentatively suggest that secondary representation is not merely a lower bound on basic moral agency, but that its presence is a strong indicator of such.

We can say something similar about metarepresentation. It is not present in animals, or in children under the age of about 3.5 years. Given that moral judgement, and therefore flexible moral agency, requires metarepresentation, this implies that it is restricted to human beings above this age. But the same factors suggesting that secondary representation is a strong indicator of basic moral agency suggest the same with respect to metarepresentation and flexible moral agency.

[104] Whiten and Suddendorf (2001) have explicitly claimed this is true of great apes, while Perner (1991) and Doherty (2007) have made the same claim about toddlers.
[105] See, for instance, Dennett (1987).

CONCLUSION

As discussed in Chapter One, moral agency is characterised by the possession of certain abilities, which enable certain actions, which when performed open agents up to certain responses. There are many ways of filling in the relevant details, depending on which abilities, actions, and responses one takes to be important. Almost any account of moral responsibility could be repurposed as an account of moral agency, provided that its responsibility conditions were agential abilities. Nonetheless, my primary aim was to devise two such accounts for two specific purposes.

In Chapter One, I developed the first of these accounts, *basic moral agency*. My purpose here was to be maximally inclusive, such that any *other* plausible account of moral agency would likewise exclude any being that failed to be a basic moral agent. I developed the account by first focusing on the specific responses of praise and blame, understood as evaluative attitudes, because these responses are the foundation for other responsibility practices, such as reward and punishment. Having done this, I found the most plausible justification for these responses in the work of Nomy Arpaly, whose account of moral responsibility holds that agents are praiseworthy or blameworthy depending on their moral motivation. In short, praiseworthy agents are those who act rightly out of good will, while blameworthy agents are those who act wrongly, either out of ill will or out of insufficient good will.

I extended this account by giving an analysis of good and ill will, according to which they are characterised by desires with specific representational content. I also distinguished blameworthy acts of insufficient good will from blameless acts according to whether agents are capable of having desires with this content. The ability to have such desires thus serves as a demarcation criterion between basic moral agents and beings who are not moral agents.

I claimed that this representational content must be pluralist with respect to standard theories of normative ethics, such that it does not exclude any being as a moral agent for failing to be motivated by reasons specific to particular normative theories. To that end, I argued that moral agents must be motivated by concern for others' welfare, respect for their autonomy, or by considerations of fairness. I argued that in all three cases, it is impossible to be appropriately motivated without the ability to have desires about others' mental states.

In Chapter Two, I considered other responsibility conditions that often appear to be necessary for moral agency, but which are not necessary for basic moral agency in the sense developed in the previous chapter. I began with the historical condition, according to which agents are only responsible for their behaviour if they have a specific causal history. I argued that while this may be a necessary condition for the justification of punishment, it is not necessary for the justification of blame.

I then turned to the epistemic condition, according to which agents are only responsible for their behaviour if they could have known whether their action was wrong or right. I observed that while this has some intuitive force with respect to ignorant wrongdoing, our intuitions with respect to ignorant rightdoing seem to pull in the opposite direction. Agents seem to be praiseworthy for doing the right thing out of good will, even when they believe themselves to be in the wrong. Regarding cases of ignorant wrongdoing, I observed that

the intuitive force of these cases seemed to be much stronger in cases of insufficient good will than in cases of ill will. In cases of ill will, I observed that agents can intentionally act wrongly without knowingly acting wrongly. By contrast, agents' blameworthiness in cases of insufficient good will is not grounded in their intentions but in their ability to have done better. If they could not have done better by virtue of their ignorance, then they are not blameworthy for their wrongdoing. Thus, I argued that the epistemic condition has a limited role to play in cases of insufficient good will but that is not relevant to cases of good will and ill will.

I then turned to the endorsement condition, according to which agents are only responsible for their behaviour if they take a particular attitude toward their action, such as endorsement, identification, or ownership. I argued that this condition derives much of its intuitive force from the phenomenology of alienated actions, and that the mere experience of an action as outside one's control does not make it so.

I then turned to the control condition, according to which agents are only responsible for their actions if they exercise the right kind of control over their actions. I distinguished three types of control failure – manipulation, compulsion, and reflexes – and argued that these do not undermine responsibility except insofar as they undermine agents' abilities to act out of good or ill will.

I argued that manipulation cases derive their intuitive force from several factors, some of which are shared with the other responsibility conditions, but one of which is the fact that manipulation involves external forces acting upon agents. I argued that in this respect, manipulation is not relevantly different from causal determinism and that if one accepts standard compatibilism then one should accept an analogous compatibilism between responsibility and manipulation.

Regarding compulsion, I argued that this does not undermine responsibility for acts of good or ill will, since such acts, even when compelled, are still acts of good or ill will. I argued that this is not the case for blameworthy acts of insufficient good will, as these require agents to have been able to do otherwise. Whether this is actually the case for any particular act of insufficient good will depends on the strength of the relevant compulsion. Thus, there are at least some cases in which compulsion does undermine responsibility, but these exclude all acts of good and ill will, and those acts of insufficient good will wherein the compulsion is resistible.

Regarding reflexes, I observed that since these are not caused by desires, they do not express agents' quality of will and thereby do not render agents responsible.

The upshot of Chapter Two is that the historical, epistemic, endorsement, and control conditions are all unnecessary for basic moral agency, although one or more of these conditions may be necessary for more restrictive accounts of moral agency, such as accounts that aim to determine which agents are appropriate targets of punishment. Having argued this, I tentatively conclude that the condition discussed in Chapter One – the ability to have desires about others' mental states – is not only necessary for basic moral agency but also sufficient for it. Nonetheless, this conclusion remains tentative because while the four conditions discussed in Chapter Two are the most prominent conditions in the moral responsibility literature, it is conceivable that other conditions are necessary for moral agency and I have not ruled out such conditions completely.

In Chapter Three, I shifted my focus away from basic moral agency toward two other practices in our moral lives: justification and moral improvement. While these practices

are not necessary for moral praise and blame, they play an important role in our moral lives and the type of moral agency that enables these practices is worthy of analysis in its own right. To that end, I developed a second account of moral agency: *flexible moral agency*.

I began with a distinction between the role of heat in moving both hot air balloons and heat guided missiles, and observed that the difference is that the former responds directly and inflexibly to heat, whereas the latter responds flexibly to a representation of heat. I claimed that this is a case of a more general distinction between direct causation and guidance, such that guidance in general involves the use of representations.

Applied to moral agency, this distinction denotes the difference between moral motivation, as exhibited by basic moral agents, and moral guidance, such that moral agents can respond flexibly to moral reasons by using representations of these reasons. However, I observed that unlike the difference between the hot air balloon and the heat guided missile, the difference between moral motivation and moral guidance cannot be solely due to the use of representation in guidance, since moral motivation also relies on representation in the form of desires about others' mental states.

Instead, I argued that the relevant difference between moral motivation and moral guidance is that the latter involves representation not only of the right making features of the act, but of these features *as* right making features. For instance, while moral motivation might involve a desire to prevent suffering without necessarily representing suffering as bad, an analogous case of moral guidance would involve representing the badness of suffering. I observed that agents respond flexibly to representations of badness and other evaluative properties by forming moral judgements, which are best understood as beliefs rather than desires. Finally, I observed that while different kinds of evaluative properties exist, the relevant property for guidance is *desirability*, such that moral judgements can inform our actions only if they represent what we ought to be motivated to do.

Following this analysis of moral guidance as moral judgements about the desirability of actions, I considered the role of moral guidance in our practices of justification, excuse, and apology, and in various methods of moral improvement.

I argued that justification, whether to oneself or to others, necessarily involves moral guidance, as it involves the communication of moral reasons.

Excuses, by contrast, need not straightforwardly involve the use of moral reasons in this way. People typically make excuses to deflect blame, and in doing so, they generally have some understanding of the difference between blameworthy and non-blameworthy behaviour. While this understanding typically involves moral judgements, this need not always be the case. It is conceivable that a child could be aware that some actions invite blame and that others do not without understanding the grounds for this distinction. Moreover, it is conceivable that a child could offer an excuse as a conditioned response to the experience of being blamed. In neither case is the child acting for moral reasons, but both situations presuppose background conditions involving the use of moral reasons, such as a general practice of offering excuses in order to deny blameworthiness.

Apologies, like excuses, need not straightforwardly involve the use of moral reasons. Unlike excuses, apologies don't function to deflect blame but to repair relationships. Typically, they do this by acknowledging wrongdoing or the perception of such. Such cases often do involve moral judgements about which actions are wrong, or perceived as wrong by the recipient of the apology. Even non-genuine apologies of the form "I'm sorry you feel that way" acknowledge the badness of hurting others' feelings, and thereby involve the use of moral judgements. However, as with excuses, it is conceivable that one may apologise either without understanding the grounds for one's behaviour or as a

conditioned response to previous apologies. In neither case are moral judgements the direct cause of these apologies, although again, such apologies do seem to require background conditions under which apologies are generally given in response to moral judgments of wrongdoing.

Having established that justification necessarily involves moral guidance, I then showed that justification plays an important role in the practice of moral improvement, specifically the improvement of one's own moral character and that of others. This is most clearly evident when one directly explains why particular actions are right or wrong. This is an effective strategy for improving moral character, insofar as the recipient internalises and begins to act upon the relevant moral reasons.

I argued that this is likewise the case for two other common methods of moral improvement: role modelling and discipline. While neither method *requires* the use of justification, reliable improvement of moral character is unlikely to occur unless one of the two agents involved uses justification.

I argued that these two abilities, the ability to justify our actions and the ability to improve our moral character, impose a normative requirement upon flexible moral agents to do so. Given that justification and improvements in moral character are more likely to lead to morally right action, flexible moral agents have a moral duty to engage in these practices.

In Chapter Four, I considered the empirical question of which agents are moral agents. I began with Josef Perner's distinction between primary representation, secondary representation, and metarepresentation, which has proven useful in explaining the different abilities of babies, toddlers, and pre-school children. I argued that despite its explanatory power, this distinction is built on a conceptual misunderstanding of the functions of representation, specifically that representation primarily functions to represent the world as it is and only secondarily functions to represent nonreal situations. Given that beliefs and desires are interdependently necessary for intentional action, the function of representing nonreal situations, as is the case for desires, cannot be secondary to the function of representing real situations.

Because Perner initially conceived of the distinction between primary and secondary representation as depending on the primary and secondary functions of representation, this posed a problem. Nonetheless, I argued for reconceiving this distinction as depending on *content availability*. On this way of thinking, primary representation involves two discrete models: a belief-like model representing the real world and a desire-like model representing goal states. These models are insulated from one another, such that representational content from one model cannot be imported into the other. Secondary representation involves multiple models, including in addition to models of the real world and of goal states, models of hypothetical scenarios and of real scenarios from other perspectives. Nonetheless, it is not characterised by the number of models but instead by the ability of these models to draw on representational content from other models. Finally, metarepresentation adds an additional type of model: those that function to represent other models. Despite these changes to Perner's framework, this three-part distinction between primary, secondary, and metarepresentation remains useful for distinguishing between the abilities of babies, toddlers, and pre-school children.

Given that both basic and flexible moral agency are characterised by the representation of mental states, this framework made it possible to identify the agents with the representational abilities necessary for moral agency. In the case of flexible moral agency, I argued that metarepresentation is necessary. This is because moral judgements about the

desirability of situations necessarily involve the representation of these situations as the content of a desire, albeit a generalised one.

In the case of basic moral agency, I argued that metarepresentation is necessary to respond to concerns of autonomy or fairness, but that it is not always necessary to respond to concerns of welfare, specifically when one responds to experiential states, such as pleasure or suffering. Although there is disagreement between philosophers as to whether these states are representational, I argued that responding to concerns of welfare need not involve responding to their representational aspects; a desire to prevent pain, for instance, need not represent this pain as a representation of bodily damage, but instead may merely represent pain as unpleasant.

Having established that metarepresentation is not necessary for basic moral agency, I then argued that secondary representation is necessary on the grounds that representation of others' experiences involves the correct kind of content availability. Specifically, content from a representational model of another person's experiences is available for use in the agent's model of their own goals.

Given that basic moral agency requires secondary representation, while flexible moral agency requires metarepresentation, I then surveyed the empirical literature to determine the age at which these develop and whether they are present in any nonhuman animals. To do this, I considered four abilities: pretend play, mirror self-recognition, theory of mind, and inhibitory executive function. The empirical evidence largely substantiates the claims that the former two abilities are indicative of secondary representation and develop at around 15-18 months of age, while the latter two abilities are indicative of metarepresentation and develop at around 3.5-4 years of age. This puts the lower bound for basic moral agency at 15 months and the lower bound for flexible moral agency at 3.5 years, although the evidence shows that the development of metarepresentation, and therefore of flexible moral agency, may be delayed if the child is autistic or has difficulties with language acquisition, as is sometimes the case for deaf children of hearing parents.

Regarding non-human animals, evidence from empirical studies on mirror-self recognition and theory of mind strongly suggests that no non-human species are capable of metarepresentation, and that secondary representation is limited to great apes, dolphins, and elephants. This suggests these species, but only these species, could be basic moral agents. It also suggests that flexible moral agency is limited to human beings. The idea that some animals may be morally praiseworthy or blameworthy for their actions is counterintuitive, but it is a potential implication of my account of moral agency, and one that deserves further study.

Finally, I tentatively claimed that the development of secondary representation in toddlers, apes, dolphins, and elephants, as well as the development of metarepresentation in 3.5-year-old children, not only mark the lower bound for basic and flexible moral agency, respectively, but that moral agency of both types develops quite soon after the development of the respective representational ability. That is, once a child develops secondary representation, it will only be a short time before he develops basic moral agency, and likewise for the development of metarepresentation and flexible moral agency. In general, the abilities associated with each stage of representation tend to develop within a short time of each other. Mirror self-recognition, pretend play, emotional meltdowns, and the desire to do things for oneself, for instance, all develop within a few months. Moral motivation is another manifestation of this same general ability, and I would be surprised if it took significantly longer to develop. I suspect the same is true of

metarepresentation and moral judgement, for much the same reasons: its associated abilities, theory of mind and inhibitory executive function, develop within a few months of each other.

Nonetheless, this remains a tentative claim because while there is a good deal of empirical research on the abilities associated with secondary and metarepresentation, I am unaware of research on the specific abilities that I take to be constitutive of moral motivation and moral judgement. For instance, if the empirical work were to show that children begin to have desires about others' mental states at around 18 months, then I would be more confident in this claim. Likewise, if the empirical work showed that children begin to have beliefs about the desirability of situations at 3.5 years. At this stage, I can only be confident in claiming that moral motivation and moral judgement do not develop before these ages, respectively. Determining these ages more precisely would require doing the relevant empirical research.

In conclusion, I hope to have given a plausible account of the kinds of moral agency that underpin our practices of moral evaluation and our ability to justify our actions and to improve our character, as well as a guide to where to find these kinds of moral agency. Having done so, it is my hope that we can use this guide to properly identify moral agents of both types in the real world.

REFERENCES

Allen, C., (1999). Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis*, 51(1), p.537.

Alvarez, M., (2017). Reasons for Action: Justification, Motivation, Explanation, *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*, Zalta, E.N. (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/>.

Amsterdam, B., (1972). Mirror self-image reactions before age two. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 5(4), p.297.

Anderson, J.R., (1984). The development of self-recognition: A review. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 17(1), p.35.

Anscombe, G.E.M., (1957). *Intention.* Blackwell.

Apperly, I.A. & Butterfill, S.A., (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological Review*, 116(4), p.953.

Aristotle, (350BCE/1998). *The Nicomachean Ethics*. Dover Publications.

Arpaly, N., (2002). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press.

Arpaly, N., (2004). Which Autonomy?. In O.'Rourke, M. & Campbell, J.K., (eds.), *Freedom and Determinism*. MIT Press.

Arpaly, N., (2005). How it is not "just like diabetes": Mental disorders and the moral psychologist. *Philosophical Issues*, 15(1), p.282.

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

Asendorpf, J.B., Warkentin, V., & Baudonniere, P.M., (1996). Self-awareness and other-awareness II: Mirror self-recognition, social contingency awareness, and synchronic imitation. *Developmental Psychology*, 32(2), p.313.

Austin, J. L., (1956). Ifs and cans. Reprinted in Berofsky, B. (ed.), (1966). *Free Will and Determinism*. Harper & Row.

Baillargeon, R., Buttelmann, D., & Southgate, V., (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, 46, p.112.

Baron, M.W. (1997) *Kantian Ethics*. In Baron, M.W., Pettit, P., & Slote, M., (1997). *Three Methods of Ethics: A Debate*. Wiley-Blackwell.

Baron, M.W., Pettit, P., & Slote, M., (1997). *Three Methods of Ethics: A Debate*. Wiley-Blackwell.

Baron-Cohen, S., Leslie, A.M., & Frith, U., (1985). Does the autistic child have a "theory of mind"?. *Cognition*, 21(1), p.37.

Barone, P., Corradi, G., & Gomila, A., (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57, p.101350.

Bennett, C., (2008). *The Apology Ritual: A Philosophical Theory of Punishment*. Cambridge University Press.

Bentham, J., (1780/2007). *An Introduction to the Principles of Morals and Legislation*. Dover Publications.

Black, S. & Tweedale, J., (2002). Responsibility and alternative possibilities: The use and abuse of examples. *The Journal of Ethics*, 6(3), p.281.

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 18(2), p.227.

Bornstein, M.H., Arterberry, M.E., & Lamb, M.E., (2013). *Development in infancy: A contemporary introduction*. Psychology Press.

Broome, J., (1991). *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

Byrne, R.W. & Whiten, A., (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.

Call, J. & Tomasello, M., (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, 70(2), p.381.

Carlson, S.M., Claxton, L.J., & Moses, L.J., (2015). The relation between executive function and theory of mind is more than skin deep. *Journal of Cognition and Development*, 16(1), p.186.

Carruthers, P., (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Carruthers, P., (2019). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press.

Chalmers, D.J., (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chisholm, R.M., (1964). Human Freedom and the Self. Reprinted in Kane, R. (ed.) (2001), *Free Will*. Blackwell.

Churchland, P.M., (1988). Folk psychology and the explanation of human behavior. *Philosophical Perspectives*, 3, p.225.

Clarke, R., (1993). Towards a Credible Agent-Causal Account of Free Will. *Noûs*, 27(2), p.191.

Crane, T., (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. Routledge.

Crisp, R., (2006). *Reasons and the Good*. Clarendon Press.

Cullity, G.M., (2018). *Concern, Respect, and Cooperation*. Oxford University Press.

Damasio, A.R., (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Davidson, D., (1963), Actions, Reasons, and Causes. Reprinted in Berofsky, B., (ed.) (1966), *Free Will and Determinism*. Harper & Row.

Davidson, D., (1982). Rational animals. *Dialectica*, 36(4), p.317.

Dennett, D.C., (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1(4), p.568.

Dennett, D.C., (1987). *The Intentional Stance*. MIT Press.

Dennett, D.C., (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.

de Waal, F.B., (2011). *Moral Behavior in Animals*. URL = <https://www.ted.com/talks/frans_de_waal_moral_behavior_in_animals>.

de Waal, F.B., Leimgruber, K., & Greenberg, A.R., (2008). Giving is self-rewarding for monkeys. *Proceedings of the National Academy of Sciences*, 105(36), p.13685.

Doherty, M., (2008). *Theory of mind: How children understand others' thoughts and feelings*. Psychology Press.

Dretske, F., (1988). *Explaining Behavior*, MIT Press.

Dunbar, R.I.M., (1998). *Grooming, gossip, and the evolution of language*. Harvard University Press.

Eshleman, A., (2019). Moral Responsibility, *The Stanford Encyclopedia of Philosophy (Fall 2019 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/moral-responsibility/>.

Eyal, N., (2019). Informed Consent, *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/informed-consent/>.

Fischer, J.M., (1987). Responsiveness and moral responsibility. Reprinted in Fischer, J.M., (2006), *My Way: Essays on Moral Responsibility*. Oxford University Press.

Fischer, J.M., (1997). Responsibility, control, and omissions. Reprinted in Fischer, J.M., (2006), *My Way: Essays on Moral Responsibility*. Oxford University Press.

Fischer, J.M., (2004). Free will and moral responsibility, Reprinted in Fischer, J.M., (2006), *My Way: Essays on Moral Responsibility*. Oxford University Press.

Fischer, J.M., Kane, R., Pereboom, D., & Vargas, M., (2007). *Four Views on Free Will*. Wiley-Blackwell.

Foot, P., (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, 5, p.5.

Frankfurt, H.G., (1969). Alternate Possibilities and Moral Responsibility. Reprinted in Frankfurt, H.G., (1988), *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Frankfurt, H.G., (1971). Freedom of the will and the concept of a person. Reprinted in Frankfurt, H.G., (1988), *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Frankfurt, H.G., (1977). Identification and externality. Reprinted in Frankfurt, H.G., (1988), *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Frankfurt, H.G., (1987). Identification and Wholeheartedness. Reprinted in Frankfurt, H.G., (1988), *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Gallup, G., (1970). Chimpanzees: self-recognition. *Science*, 167(3914), p.86.

Geppert, U. & Küster, U., (1983). The emergence of 'wanting to do it oneself': A precursor of achievement motivation. *International Journal of Behavioral Development*, 6(3), p.355.

Gert, B. & Gert, J., (2020). The Definition of Morality, *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/morality-definition/>.

Gopnik, A. & Astington, J.W., (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 59(1), p.26.

Griffin, J., (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press.

Hacker, P., (2007). *Human Nature: The Categorical Framework*. Blackwell.

Haidt, J., (2010). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hare, R.M., (1981). *Moral thinking: Its levels, method, and point*. Oxford University Press.

Heathwood, C., (2014). Subjective Theories of Well-Being. In Eggleston, B., & Miller, D., (eds.), *The Cambridge Companion to Utilitarianism*. Cambridge University Press.

Heyes, C., (2014). False belief in infancy: A fresh look. *Developmental Science*, 17(5), p.647.

Hobart, R.E., (1934). Free will as involving determination and inconceivable without it. Reprinted in Berofsky, B. (ed.) (1966), *Free Will and Determinism*. Harper & Row.

Hobbes, T. (1651/2017). *Leviathan*. Penguin.

Hogrefe, G.J., Wimmer, H., and Perner, J., (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 57(3), p.567.

Hume, D. (1738/1985). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*. Penguin.

Ichikawa, J.J., and Steup, M., (2018). The Analysis of Knowledge, *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*, Zalta, E.N. (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

Isaacs, T., (2011). *Moral responsibility in collective contexts*. Oxford University Press.

Jones, L.B., Rothbart, M.K., and Posner, M.I., (2003). Development of executive attention in preschool children. *Developmental Science*, 6(5), p.498.

Joyce, R., (1999). Apologizing. *Public Affairs Quarterly*, 13(2), p.159.

Joyce, R., (2005). *The Evolution of Morality*. Bradford.

Kane, R. (2007). *Libertarianism*. In Fischer, J.M., Kane, R., Pereboom, D., & Vargas, M., (2007). *Four Views on Free Will*. Wiley-Blackwell.

Kant, I., (1785/2002). *Groundwork for the Metaphysics of Morals*. Oxford University Press.

Kaplan, G., (2019). *Bird Bonds*. Macmillan.

Kennett, J., (2001). *Agency and Responsibility: A Common-Sense Moral Psychology*. Oxford University Press.

Kennett, J. & Fine, C., (2008). Internalism and the evidence from psychopaths and "acquired sociopaths.". In Sinnott-Armstrong, W.E., (ed.), *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development*. MIT Press.

Kennett, J. & Smith, M., (1996). Frog and toad lose control. *Analysis*, 56(2), p.63.

Kirk, R., (1974). Zombies v. Materialists. *Aristotelian Society*, Supplementary Volume 48(1), p.135.

Kirk, R., (2021). Zombies, *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/zombies/>.

Korsgaard, C.M., (1997). The Normativity of Instrumental Reason. In Cullity, G.M. & Gaut, B., (eds.), *Ethics and Practical Reason*, Oxford University Press.

Korsgaard, C.M., (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Levy, N., (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.

Lieberman, A.F., (2017). *The emotional life of the toddler*. Simon & Schuster.

List, C. & Pettit, P., (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Mackie, J.L., (1977). *Ethics: Inventing Right and Wrong*. Penguin.

McCloskey, H.J., (1957). An examination of restricted utilitarianism. *The Philosophical Review*, 66(4), p.466.

McDowell, J., (1995). Might There Be External Reasons? In Altham, J.E.J. & Harrison, R., (eds.), *World, Mind and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, Cambridge University Press.

McGeer, V., (2008). Varieties of moral agency: Lessons from autism (and psychopathy). In Sinnott-Armstrong, W.E., (ed.), *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development*. MIT Press.

McKenna, M., (2005). Reasons reactivity and incompatibilist intuitions. *Philosophical Explorations*, 8(2), p.131.

McKenna, M. & Coates, D.J., (2021). Compatibilism, *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*, Zalta, E.N. (ed.), URL = <https://plato.stanford.edu/archives/fall2021/entries/compatibilism/>.

McMahan, J., (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.

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

Mele, A.R., (2012). *Backsliding: Understanding Weakness of Will*. Oxford University Press.

Mill, J.S., (1863/2002). *Utilitarianism*. Hackett Publishing.

Moore, C., (2010). Understanding Self and Others in the Second Year. In Brownell, C.A. & Kopp, C.B., (eds.), *Socioemotional development in the toddler years: Transitions and transformations*. Guilford Press.

Moses, L.J. & Flavell, J.H., (1990). Inferring false beliefs from actions and reactions. *Child Development*, 61(4), p.929.

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

Nielsen, M. & Dissanayake, C., (2004). Pretend play, mirror self-recognition and imitation: A longitudinal investigation through the second year. *Infant Behavior and Development*, 27(3), pp.342-365.

Nielsen, M., Dissanayake, C., & Kashima, Y., (2003). A longitudinal investigation of self–other discrimination and the emergence of mirror self-recognition. *Infant Behavior and Development*, 26(2), p.213.

Nozick, R., (1974). *Anarchy, State, and Utopia*. Basic Books.

Onishi, K.H. & Baillargeon, R., (2005). Do 15-month-old infants understand false beliefs?. *Science*, 308(5719), p.255.

Parfit, D., (1997). Reasons and Motivation, *Proceedings of the Aristotelian Society*, Supplementary Volume 71, p.99.

Parfit, D., (2011). *On What Matters: Two-Volume Set*. Oxford University Press.

Pavlov, I.P. & Anrep, G.V., (1928). Conditioned Reflexes. *Journal of Philosophical Studies*, 3(11), p.380.

Pearce, J.M., (2013). *Animal learning and cognition: an introduction*. Psychology Press.

Pereboom, D. (2007). *Hard Incompatibilism*. In Fischer, J.M., Kane, R., Pereboom, D., & Vargas, M., (2007). *Four Views on Free Will*. Wiley-Blackwell.

Perner, J., (1991). *Understanding the representational mind*. MIT Press.

Perner, J., (2014). Commentary on Ted Ruffman's "Belief or not belief:…". *Developmental Review*, 34(3), p.294.

Perner, J., Leekam, S.R., & Wimmer, H., (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2), p.125.

Perner, J. & Ruffman, T., (2005). Infants' insight into the mind: How deep?. *Science*, 308(5719), p.214.

Peterson, C.C. & Siegal, M., (1999). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science*, 10(2), p.126.

Pettit, P. & Smith, M., (1990). Backgrounding desire. *Philosophical Review*, 99 (4):565-592.

Premack, D. & Woodruff, G., (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4(4), p.515.

Rawls, J., (1971/1999). *A Theory of Justice*. Belknap Press.

Regan, T., (1983). *The Case for Animal Rights*. University of California Press.

Reiss, D. & Morrison, R., (2017). Reflecting on mirror self-recognition: A comparative view. In. Call, J., Burghardt, G.M., Pepperberg, I.M., Snowdon, C.T., & Zentall T., (eds.), *APA handbook of comparative psychology: Perception, learning, and cognition*. American Psychological Association.

Richerson, P.J. & Boyd, R., (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago University Press.

Rosati, C.S., (2016). Moral Motivation, *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, Zalta, E.N. (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/moral-motivation/>.

Rosen, G., (2002). The Case for Incompatibilism, *Philosophy and Phenomenological Research*, 64(3), p.699.

Rowlands, M., (2015). *Can animals be moral?*. Oxford University Press.

Ruben, D.H., (2003). *Explaining explanation*. Routledge.

Ruffman, T., (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, 34(3), p.265.

Russell, P., 2010. Selective hard compatibilism. In Campbell, J.K, O'Rourke, M., & Silverstein, H.S., (eds.), *Action, ethics, and responsibility*. MIT Press.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T., (1991). The 'windows task as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, 9(2), p.331.

Salmon, W.C., (1989). Four Decades of Scientific Explanation. *Minnesota Studies in the Philosophy of Science*, 13, p.3.

Scanlon, T.M., (1998). *What We Owe to Each Other*. Belknap Press.

Searle, J., (1983). *Intentionality*. Cambridge University Press.

Schick, B., De Villiers, P., De Villiers, J., & Hoffmeister, R., (2007). Language and theory of mind: A study of deaf children. *Child Development*, 78(2), pp.376-396.

Schlick, M., (1963). Problems of Ethics. Reprinted in Berofsky, B., (ed.) (1966), *Free Will and Determinism*. Harper & Row.

Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.

Schroeder, T., (2020). Desire, *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*, Zalta E.N., (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/desire/>.

Schwitzgebel, E., (1999). Representation and desire: A philosophical error with consequences for theory-of-mind research. *Philosophical Psychology*, 12(2), p.157.

Schwitzgebel, E., (2017). An Argument Against Every General Theory of Consciousness, *The Splintered Mind*, URL = <https://schwitzsplinters.blogspot.com/2018/05/an-argument-against-every-single.html>.

Shafer-Landau, R., (1998). Moral Motivation and Moral Judgment, *Philosophical Quarterly*, 48, p.353.

Sher, G., (2009). *Who Knew?: Responsibility Without Awareness*. Oxford University Press.

Sinhababu, N., (2017). *Humean Nature*. Oxford University Press.

Singer, P., (1979). *Practical Ethics*. Cambridge University Press.

Smart, J.J.C. & Williams, B., (1973). *Utilitarianism: For and Against*. Cambridge University Press.

Smith, M., (1994). *The Moral Problem*. Blackwell.

Sobel, D. & Copp, D., (2001). Against direction of fit accounts of belief and desire. *Analysis*, 61(1), p.44.

Sober, E. & Wilson, D.S., (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Sroufe, L.A., (1997). *Emotional development: The organization of emotional life in the early years*. Cambridge University Press.

Strawson, G.J., (1994). The impossibility of moral responsibility. *Philosophical Studies*, 75(1-2), p.5.

Strawson, P.F. (1962). Freedom and Resentment. Reprinted in Fischer, J.M. & Ravizza, M., (eds.) (1993), *Perspectives on Moral Responsibility*. Cornell University Press.

Stump, E., (1993). Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will. In Fischer, J.M. & Ravizza, M., (eds.), *Perspectives on Moral Responsibility*. Cornell University Press.

Suddendorf, T. & Whiten, A., (2001). Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin*, 127(5), p.629.

Tognazzini, N. & Coates, D.J., (2021). Blame, *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/blame/>.

Tomasello, M. & Call, J., (1997). *Primate cognition*. Oxford University Press.

Tooley, M., (1972). Abortion and infanticide. *Philosophy and Public Affairs*, 2(1), p.37.

van Gulick, R., (2021). Consciousness, *The Stanford Encyclopedia of Philosophy (Winter 2021 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/consciousness/>.

van Inwagen, P., (1975). The incompatibility of free will and determinism. Reprinted in Kane, R. (ed.) (2001), *Free Will*. Blackwell.

van Roojen, M., (2018). Moral Cognitivism vs. Non-Cognitivism, *The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/moral-cognitivism/>.

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

Walen, A., (2021). Retributive Justice, *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*, Zalta, E.N., (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/>.

Wallace, R.J., (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Watson, G., (1975). Free agency. Reprinted in Watson, G., (2004), *Agency and Answerability: Selected Essays*. Oxford University Press.

Watson, G., (1996). Two faces of responsibility. Reprinted in Watson, G., (2004), *Agency and Answerability: Selected Essays*. Oxford University Press.

Watson, G., (2002). Volitional Necessities. Reprinted in Watson, G., (2004), *Agency and Answerability: Selected Essays*. Oxford University Press.

Wellman, H.M., Cross, D., & Watson, J., (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3), p.655.

Whiten, A. & Suddendorf, T., (2001). Meta-representation and secondary representation. *Trends in Cognitive Sciences*, 5(9), p.378.

Wimmer, H. & Perner, J., (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), p.103.

Wolf, S., (1987). Sanity and the Metaphysics of Responsibility. Reprinted in Kane, R. (ed.) (2001), *Free Will*. Blackwell.

Wolf, S., (1990). *Freedom Within Reason*. Oxford University Press.

Wood, A.W., (2007). *Kantian Ethics*. Cambridge University Press.