What Makes Us Responsible: Fischer and Ravizza on Implicit Attitudes

Michael Ledger
Bachelor of Arts – Psychology

In partial fulfilment of the degree Master of Research (Philosophy)

Macquarie University
Philosophy Department

9/10/2015

Table of Contents

Abstract

Implicit attitudes are a mental construct which can affect our behaviours. However, the influence of these attitudes can sometimes lead to their being manifested in potentially prejudicial or discriminatory behaviours. Because of the potentially harmful effects of these behaviours, it is important that theories of moral responsibility be able to properly regard behaviours which manifest our implicit attitudes. However, there is considerable divergence in opinions in the literature surrounding the nature of implicit attitudes and the effects they have on behaviour. Therefore, determining the extent to which we may be held morally responsible for our actions manifesting our implicit attitudes can prove difficult.

The theory of moral responsibility put forward by Fischer and Ravizza (1998) presents a criterion for responsibility which I argue is capable of sufficiently responding to the challenges posed by implicit attitudes. The following project is therefore an examination of the Fischer and Ravizza theory, and its notion of a reasons-responsive deliberative mechanism. This project is also an inquiry into the nature of implicit attitudes and the role they play in this deliberative mechanism. Thus I ask the question, 'according to a Fischer and Ravizza theory of responsibility, to what extent may we be held morally responsible for our actions manifesting our implicit attitudes'.

Statement of Candidate

I, Michael Ledger, certify that the work contained in this project, titled, ‘What Makes Us Responsible: Fischer and Ravizza on Implicit Attitudes’, has not been previously submitted for a higher degree to any other university or institution, nor has it been previously submitted as a component for a higher degree to any other university or institution.

I, Michael Ledger, certify that this project is an original piece of work, and that any other sources of information, or assistance utilised in the development of this work, have been appropriately acknowledged throughout.

Michael Ledger

Acknowledgments

When undertaking a project of this magnitude, surrounding oneself with the right support, wisdom, and knowledge is paramount. In light of this, there are a number of parties I wish to acknowledge, and to whom I wish to express my eternal gratitude.

First and foremost, I would like to thank my supervisor, Neil Levy, for his extensive contributions to this project. Truly, without your constant feedback, your ever-willingness to discuss and ponder ideas, and your patience with me through the trying times, this project would not have reached its completion. To say that working with you was an insightful experience would be a severe understatement. I have learnt much from working with you, and I am very much grateful for having had the opportunity to do so. Thank you, Neil, for all your help.

I also owe thanks to my associate supervisor, Colin Klein. Your support in the early days of this project were surely paramount to its fruition. More so than this, your friendly and cheerful optimism served to help me believe in myself and what I could accomplish. Thank you Colin for your tireless help and encouragement.

I wish to extend thanks to Jeanette Kennett, for her considerable contributions to the MRes program as a whole, and her tireless efforts to keep things running smoothly. I also want to give special thanks to my fellow MRes students, especially to Jonny and Peter, for their support, encouragement and friendship over the last two years. Whether you both know it or not, it has been a privilege to undertake this journey with you, and I look up to you both, and hold you both in the highest regard. Together, you two motivated me to try my very best. Thank you so much for that.

Of course, I very much thank my family for their unwavering patience, understanding, encouragement, and support during this time; Paul, Marie, and Chris. As I said, surrounding

oneself with the right support is paramount when undertaking a project such as this, and I could not have asked for a more remarkable and loving bunch of people in my corner than you.

I want to extend my thanks to those friends of mine who ensured that I did not lose myself in my papers and books throughout the year, and reminded me to always keep one foot in the real world. Patrick, Kimberley, Nicholas, Jessica-Rose, and Courtney, I have you all to thank for ensuring I maintained my sanity through some pretty arduous months. Your love and support has not gone unnoticed.

Finally, to Arielle. Words cannot express my gratitude for all that you have done. You were there with me through the highs, the lows, the smiles and the tears. You did not let me give up on myself and you never once stopped believing in me. Thank you.

What Makes Us Responsible: Fischer and Ravizza on Implicit Attitudes

Introduction

Many different moral theories have been constructed over the centuries, each endeavouring to explain what makes us morally responsible. Each theory presents its own unique account of moral responsibility, and the account presented and advocated by Fischer and Ravizza (1998) is no different.

According to a Fischer and Ravizza (1998) framework of moral responsibility, holding someone morally responsible for their actions requires that person to have a certain type of control over their actions. While I will go in to more detail on the specifics of this account later in chapter one, Fischer and Ravizza contend we may only hold people morally responsible if they can recognise a coherent and consistent pattern of reasons for acting or not acting (including some moral reasons), and then if they possess enough control to (at least try to) respond to those reasons. It is for these kinds of reasons that we are unable to hold psychopaths, very young infants, and in some cases the elderly, responsible for their actions (Fischer and Ravizza, 1998). This is because they simply do not satisfy these conditions for moral responsibility. These subjects are not always able to recognise or respond to a consistent and coherent pattern of (moral) reasons for acting.

In each of these examples we can conceive of instances where the criminal justice system perhaps shows, for lack of a better word, a bit more ‘leniency’ in ascribing responsibility. Even our intuitions lead us to feel like these individuals aren’t responsible; at least, not to the same degree as a typically functioning agent.[1] It is accounting for exemplary cases such as these which have challenged moral theorists. There are, however, other cases which can challenge a moral theory and one of these challenges is currently a prevalent social concern; that is implicit attitudes.

An examination of the literature surrounding implicit attitudes reveals that there is significant conjecture surrounding their nature and operation.[2] Opinions across the literature diverge extensively in regards to exactly what implicit attitudes are, how they operate, what kind of control we have over them, and what effect they have over behaviour (if any). Therefore, constructing a cohesive definition or list of properties indicative of implicit attitudes is difficult at best. That being said, one can observe convergence of opinions regarding a couple of traits in particular.

A considerable amount of empirical research suggests that implicit attitudes exist and operate outside of conscious awareness, that they are automatically activated, and are introspectively inaccessible mental states. Implicit attitudes may be positive or negatively valanced attitudes which we hold in regards to objects in the world. They have also been demonstrated to at times conflict, or differ, from the content of our more explicitly expressed attitudes.[3] However, it must be remembered that even these somewhat more agreed upon characteristics are not absolute; sometimes it is the case that our implicit and explicit attitudes align with each other and no conflict is observed.

[1] I take a 'typical functioning agent' to be an agent with the appropriate array of mental, deliberative, and operational capacities. By 'typical' I mean nothing more than those faculties which one would reasonably expect to be possessed by the average person.

[2] I dedicate chapter two to the examination of this literature. As such, I will provide citations for my sources when I discuss these claims more fully in chapter two.

[3] In chapter two I elaborate further on exactly what are 'implicit' and 'explicit' attitudes and how they differ. For now, examples of explicit attitudes include the attitudes which we are able to identify consciously, or perhaps voice out loud. In other words, they are the attitudes we have which we can express 'explicitly', such as 'I believe all men and women are created equal'.

The degree to which we are consciously aware of our attitudes and their content is also a debated topic. While in some cases subjects may be unaware that they even possess implicit attitudes at all, with the right testing procedures and instruction, subjects can come to be made aware of their existence and content. However, one aspect of implicit attitudes which is particularly problematic, and raises the challenges moving in to my own project is the fact that most of the time their operation can occur outside of conscious awareness.

Suppose I were to ask if you held any prejudice against hiring women in the workplace. You argue emphatically (explicitly) that you do not; that women are just as capable as men, if not more capable, and that women deserve an 'equal shot'. Now suppose further that I ask you to select an eligible candidate for a job, and then present you with equally qualified male and female candidates (maybe even a superiorly qualified female). In a number of cases we might find you selecting the male candidate, even though there is no rational reason to hire the male candidate over the superiorly qualified female. In this instance, your selection may be guided, or otherwise influenced, by the presence some sort of implicit attitude; perhaps an implicit gender bias against women.

Examples and studies like this claim to demonstrate that implicit attitudes can affect our behaviour. However, the degree to which they may do so is still contentious; nonetheless, there is considerable evidence to support the notion that implicit attitudes do in fact play some sort of role in the processes which produce behaviour (at least in the deliberative processes involved in behaviour production).

A challenge comes in when we remember that we may not know we have these attitudes and we may not be aware of them influencing our actions. Despite this, implicit attitudes can still shape our behaviour in particular ways, and we can even manifest our implicit attitudes in our behaviour. Manifesting our implicit attitudes, as in the example above, can potentially result in prejudicial or discriminatory behaviour; alternatively, 'morally relevant behaviours'[4]. Thus some challenges arise. When our actions manifest our implicit attitudes, to what extent have our implicit attitudes contributed to the mechanism producing that behaviour? Furthermore, Fischer and Ravizza (1998) require that the mechanism producing the behaviour be, to a certain extent, responsive to reasons, in order to ascribe moral responsibility; therefore, we may also ask, to what extent might the influence of our implicit attitudes affect the degree of 'reasons-responsiveness' displayed by the mechanism? Out of this emerges another important question we must ask; to what extent are we morally responsible for our actions which manifest our implicit attitudes?

This is where my study comes in, and it has three aims. Firstly, my study seeks to provide an examination of the Fischer and Ravizza (1998) theory of moral responsibility. In the proceeding chapters, I will demonstrate that this moral theory is both realistic in the requirements it places on the moral agent, as well as being practical in its application. I will be considering as well some criticisms and challenges which have been mounted in response to this theory, and will evaluate Fischer and Ravizza's response to these queries.

Secondly, I seek to provide an inquiry in to the nature of implicit attitudes. As I have already mentioned, there is considerable tension in the literature surrounding the topic of implicit attitudes; opinions diverge and conflict, and variability is abundant. My project examines what implicit attitudes are, and what influence they have over behaviour. My inquiry in to implicit attitudes will also involve an examination of their primary and foremost testing procedures; how do they work; what can we gleam from the results; are implicit and explicit attitudes single or dual constructs. The literature surrounding implicit attitudes provides mixed responses to these

[4] In the example provided above, the job recruiter's implicit attitude in question has a negative valence; it is a negative implicit attitude in that he regards women in an unfavourable way. However implicit attitudes can also have a positive valence. For instance, I may have a particular fondness for one friend over another.

questions. Thus, one of the aims of my study is to provide a brief survey of the literature surrounding implicit attitudes and their testing procedures. I will be highlighting what I believe to be key characteristics, as well as key misconceptions and challenges about their nature and operation.

The third and primary aim of my study is that I will be asking, ‘to what extent are we morally responsible for our actions which manifest our implicit attitudes’. The aspect of this project which makes it unique, is that I will be answering this question according to the theory of moral responsibility put forward by Fischer and Ravizza (1998). While the question of moral responsibility and implicit attitudes has been asked before, little consideration has been given to how this particular theory would respond to the question.

Fischer and Ravizza (1998) put forward what has been described as “one of the most compelling accounts of moral responsibility to date” (Pereboom, 2006) because it is a very practical and plausible account of responsibility which can also be adopted by the hard determinist; hard determinism being a philosophical stand point which denies the existence of free will, but which has been known to consequently struggle with providing an account of moral responsibility. If indeed Fischer and Ravizza’s moral theory is found to be successful in its ability to provide an account of moral responsibility compatible with determinism, then that is a great boon for the determinists. If so, hard determinists might be able to mount an argument for moral responsibility in a sound and robust manner. More than this though, it is of social importance that we have an account of moral responsibility that is internally consistent, practical, and applicable to real world challenges. I will discuss this further shortly. To that end, my study looks at how Fischer and Ravizza would respond to the challenges posed by implicit attitudes. Do they do so in a satisfactory way?

Without going in to too much detail right now, Fischer and Ravizza's (1998) theory claims we can only be held morally responsible for those actions we perform which issue from a decision making process (or *deliberative mechanism*) that is moderately 'reasons responsive'. A reasons responsive deliberative mechanism is able to identify reasons to act or not act, and then act on those reasons by carrying out the appropriate action. Seeing as implicit attitudes are known to play a role in the deliberative mechanism affecting behaviour, I argue that the degree to which we can be held responsible for these behaviours is related to how reasons-responsive of the implicit attitudes. If it turns out that implicit attitudes are reasons-responsive, then we may be more eligible for ascriptions of moral responsibility. If we find, however, that implicit attitudes are not reasons-responsive, and yet still affect behaviour, then perhaps our implicitly biased job recruiter is not especially blameworthy or morally responsible. To that end, this study also aims to determine to what extent implicit attitudes are reasons responsive.

These are the goals of the current project; the questions and challenges which I aim to address. Throughout the proceeding chapters I will expand upon my research question, "according to the account of moral responsibility set out by Fischer and Ravizza (1998), to what extent are we morally responsible for our actions which manifest our implicit attitudes". However, before I do so, I wish to make some clarifying remarks.

Firstly, my project addresses the topic of moral responsibility. While I will be utilising the framework for moral responsibility set out by Fischer and Ravizza in their 1998 text, *Responsibility and Control*, I will be drawing on a definition of moral responsibility coined by Neil Levy (personal communication, April 22, 2015). On this definition, moral responsibility may entitle a person to be treated better or worse than they otherwise would have. For instance, if I am morally responsible for making a donation to a charity collection box then presumably I deserve to be treated a little better by, say, being praised or thanked for my action. I do expect there to be significant conjecture around adopting such a definition. However, the following project is an examination in to the applicability and consistency of a particular *framework* for moral responsibility, rather than a philosophical analysis of exactly what moral responsibility is itself. For the sake of my project, I find this working definition to be sufficient for my endeavours.

Secondly I wish to clarify what this project is not, and to do so I utilise a distinction drawn by Holroyd (2012). This project asks to what extent we may be held morally responsible for our actions which *manifest* our implicit attitudes. This project *does not* ask to what extent we are responsible for *having* implicit attitudes in the first place. Questions about how we come to acquire implicit attitudes, and whether we are morally responsible for possessing them, are a different issue. As such I will not be asking such questions as, 'is Mark responsible for having sexist implicit attitudes about women?'. Instead I will be asking questions along the lines of, 'is Mark responsible for his sexist implicit attitudes about women affecting his ability to fairly perform his role as a job recruiter?'. The distinction being drawn here is between having implicit attitudes in the first place, and to what extent we are responsible for being influenced by them or having them manifest in our behaviour. My project concerns the latter question and thus works off the assumption that we already have implicit attitudes, whether we are morally responsible for having them or not.[5] Again, my project concerns the behaviours which are caused by, or which manifest, our implicit attitudes.

Finally, before I proceed to chapter one, I wish to reiterate the significance of this project and express how broadly the implications of this research extend. For philosophers in particular, my study ought to provide a number of benefits. Firstly, I provide an analysis and examination in

[5] In chapter four I will argue that, once made aware of our implicit attitudes, we may be held morally responsible for failing to manage them (specifically, failing to exercise 'ecological control'). I maintain that this is still distinct from asking to what extent we are responsible for having implicit attitudes at all.

to the validity and applicability of the Fischer and Ravizza (1998) theory of responsibility in general. More so than this, my study has direct implications for the sub-discipline of metaphysics; specifically, free will debates for the deterministic position. If indeed it is the case that Fischer and Ravizza are able to provide an account of responsibility which can be adopted by the determinist, then they will have managed to provide a possible response to one of determinisms foremost criticisms; that being, how do you salvage moral responsibility if it is true that free will is incompatible with determinism? Furthermore, what is the freedom relevant condition for responsibility? Do we need to be able to 'do otherwise' in order to be held responsible for our actions?

Even simply as a theory of moral responsibility, I firmly believe this account merits considerable investigation. In bioethics, consideration of the ability to correctly recognise reasons and respond to them will affect the way we manage euthanasia and abortion debates; is the patient going through a reasonable deliberative process? In ethical theory, the inclusion of implicit attitudes will prompt us to address such questions as, 'what type of control is needed for responsibility', and 'can we be held responsible for unconscious processes'? Even in areas of agency and epistemology, what we can know and what obligations we have to discover our own implicit attitudes are questions that may be raised in light of this research.

Research such as the kind I am undertaking here is also of importance to those involved in the field of law. If theory of moral responsibility utilised in legal (and societal) settings were to adopt this approach, how might the way we handle criminal convictions and legal policies be affected? The way we manage cases of professional negligence, drink driving, even theft could be affected given the increased focus on decision making processes, and recognising and responding to reasons. In cases of criminal law, we may find that more, or perhaps fewer, people meet the criteria for being held criminally responsible. This current research of mine also raises questions of jurisprudence; how connected are moral and legal responsibility on this account? How ought the law to adapt to a changing view of moral responsibility?

Even the field of psychology is incorporated on a large scale by the inclusion of implicit attitudes. My study attempts to ascertain the nature of implicit attitudes, and how they work, through a survey of the psychological literature. This will be achieved through a survey and examination of the primary research methods used by empirical and cognitive psychologists to test implicit attitudes. Thus the results of my study could affect empirical psychology in the way we use and interpret the results from such primary testing procedures as the Implicit Association Test, or the Affect Misattribution Procedure. Results of my study will even be of combined interest in the areas of psychopathology, moral philosophy, and law, as this account of responsibility will affect the way we view the deliberative mechanisms of psychopaths and schizophrenics, and the degree to which we can hold them morally and criminally responsible.

The nature of implicit attitudes and their interaction with moral responsibility ought to earn the attention of sociologists and the general public as well. Implicit attitudes can take many forms including implicit racism, sexism, religious prejudice, and discrimination based on sexual orientation. If implicit sexism leads to increased violence against women; if implicit racism causes cultural ignorance; and if implicit religious prejudice leads to fear and terrorism, my study should prompt such questions as, how ought we respond? Can we hold these offenders responsible for their actions? Could we perhaps use this new knowledge to reduce fear in the media and increase social equality and tolerance?

I have attempted in this introduction to express the motivations and reasons for undertaking this project. I hope to have made it clear that this research of mine has broad, multi-disciplinary, and very *real* implications. This is not an armchair musing in philosophy, it is not a theoretical question for law and psychology, and these are not hypothetical benefits for sociology.

This research is both real and important for everyone. To that end, I move now to chapter one to detail and analyse the account of moral responsibility proposed by Fischer and Ravizza (1998).

Chapter One – Fischer & Ravizza

There are a number of ideas in philosophy and in our everyday intuitions about what exactly makes us morally responsible for our actions. Fischer and Ravizza (1998) present one such moral theory which is both practical and realistic, and which merits considerable philosophical examination. Throughout the proceeding chapter I will be providing an examination of Fischer and Ravizza's theory of moral responsibility, highlighting prevalent ideas and challenges.

Firstly, I will demonstrate that the idea of being able to 'do otherwise' is challenged by causal determinism. I argue that there are plausible arguments that this kind of control is impossible if causal determinism is true. I will then demonstrate how Fischer and Ravizza's (1998) distinct notions of regulative control and guidance control can be used to circumvent this well-known debate and still retain the possibility of moral responsibility.

Secondly, I will demonstrate how Fischer and Ravizza (1998) build upon their notion of guidance control in response to challenges presented by *manipulation arguments*. In doing this I will introduce and discuss Fischer and Ravizza's main and most important idea; that of *moderate reasons-responsiveness*.

Thirdly, I will review Fischer and Ravizza's (1998) theory as a whole. I will consider challenges raised by other academics and examine Fischer and Ravizza's response. I will finally argue that there are numerous strengths and benefits to the Fischer and Ravizza account, and for these reasons it merits considerable philosophical attention and examination.

When we perform an action we may intuitively say that we are morally responsible for said action because we could have instead acted some other way. For example, if I were to stop and help an elderly woman to cross the street then I might be praised for my action because I could have not helped the woman but still chose to do so. In other words, a number of moral theories, and even our intuitions, are based on the idea, or assume to some extent, the 'Principle of Alternate Possibilities'; the idea that we have the ability to act otherwise than we do, and that this is necessary in order to be responsible for our behaviour (Fischer & Ravizza, 1998). According to this, I am morally responsible for performing an action, *A*, in lieu of the fact that I have control over my actions insofar as I could have performed *not A*.

What this assumes is that there were at least two possible courses of action for me to take when I considered helping the woman, and while I chose to perform action *A*, it must have been at least possible that I could have performed action *B.*[6] Further, it is my choosing to perform action *A*, instead of action *B*, which makes me morally responsible for action *A*. However, as I mentioned earlier, causal determinism poses a challenge to this idea of control. In fact, there are plausible arguments that show this kind of control is an impossibility if we were to take causal determinism as being true.

Causal determinism is the idea that every given event is *necessitated* by a combination of antecedent events operating in tandem with the laws of nature (Hoefer, 2010). In other words, an event at time *t* necessarily occurs because the events of the past preceding time *t*, interact with the laws of nature in *a specific way*. This means that only one future is possible at any given moment; there is only one way things could happen. Even the actions of the agent are fixed; there are no options or alternatives open to him. If determinism is true, an agent can only ever act the way he does; he cannot ever 'do otherwise' because 'otherwise' does not exist.[7] Therefore, if

[6] Another way to think about this would be that there were two possible *futures* available to me, and because I chose to actualize one future and not the other, I am morally responsible for my actions.

[7] While it may be the case that an agent does not have any options or alternatives open to him at any given time, this does not equate to not having the ability to make choices. Determinism does not take away our ability to make choices. Determinism makes it such that any choice we do make is a choice we were always going to make given the events of the past leading up to that choice, and the way those events interact with the laws of nature. Even if determinism is true I can still make choices and decisions; I can deliberate and reason, and consider how I will act. But if determinism is true, then regardless of what decision I make, I was always going to make that choice. Whether or not this means that deliberation or decision-making is a pointless or redundant exercise in a

determinism is true, it would seem that it rules out the possibility of alternate possibilities. If the Principle of Alternate possibilities is in fact what makes us morally responsible, then it would seem that moral responsibility (so defined) is not compatible with determinism.[8]

Fischer and Ravizza's (1998) theory responds to this threat to moral responsibility by challenging the type of control which an agent needs to possess in order to be morally responsible. Given that Fischer and Ravizza adopt the working assumption that determinism is true, they advocate a type of control that is compatible with determinism. They do this by distinguishing between 'regulative control' and 'guidance control', and utilise the following thought experiment, 'Assassin' (Frankfurt, 1969).

Suppose Sam plans to murder the town mayor. Suppose Sam tells Jack what he is going to do. Jack also wants Sam to kill the mayor, and to ensure Sam does so, Jack implants a device into Sam's brain; should Sam decide not to kill the mayor, Jack can use his device to force Sam to kill the mayor anyway. In the end Sam kills the mayor without Jack having to intervene.

In this example, our intuitive response is still to say that Sam is morally responsible for killing the mayor, even though he could not have 'done otherwise'; either Sam would have killed the mayor himself, or Sam would have killed the mayor due to the existence of a counterfactual intervener (Jack). Fischer and Ravizza (1998) use this thought experiment to demonstrate that (at least intuitively) moral responsibility is not defined by being able to do otherwise or having alternate possibilities. A different type of control is required.

deterministic world is another issue I will not be discussing here. But suffice to say, determinism as described here does not take away our ability to make choices.

[8] This incompatibility between the Principle of Alternate Possibilities and determinism is one of the topics that is discussed in the 'free will' debate. The free will debate discusses what free will is, whether we have free will, and considers such problems as the conflict between those few compatibilist positions which accept the Principle of Alternate Possibilities, and the few determinist positions which claim we do not have free will (at least not free will as defined by the Principle of Alternate Possibilities).

Consider their (Fischer & Ravizza, 1998) second example used to distinguish between two types of control: Sally is driving her car with a driving instructor and could turn left or right. Sally turns the car to the left. When Sally does so we say that she is exercising *guidance control* over the car in lieu of the fact that Sally *controls* the car. At the same time Sally could have instead turned the car to the right. If Sally had "acted otherwise" she would have exercised *regulative control*.

Fischer and Ravizza (1998) claim an agent possesses guidance control in virtue of the fact that the agent is the one performing the action; such as Sally being the one driving. Regulative control consists in a dual power, and is the control which grants us the ability to 'do otherwise' (Fischer & Ravizza, 1998). As I have just said, guidance control may be thought of as the control we possess by being the agent performing the action[9], so exercising regulative control necessarily entails exercising guidance control as well. However it is having regulative control that entails we could have done otherwise; that there was an 'alternate sequence' of events, another possible world[10], where we would have acted otherwise. For instance, Sally might have possessed regulative control over the car because, in another world, she would have turned to the right instead of the left.

Regulative control requires us to be exercising a degree of guidance control over either course of action, but I can exercise guidance control without the possibility of doing otherwise (without regulative control). Suppose further that the driving instructor in this example has his

[9] As I go on to discuss later in the chapter (see footnote 26), there are situations where we may be the agent carrying out the action, but may not have guidance control. Cases, such as those described in manipulation arguments (for example, brainwashing, direct brain manipulation, hypnosis, etc.), are instances of carrying out an action and lacking guidance control.

[10] The notion of 'possible worlds' is a philosophical tool used to describe 'counterfactual' states of affairs; in this case, what we *would* have done in the alternate sequence is what we *did* in some other possible world. While Fischer and Ravizza utilise the philosophical tool of 'possible worlds', their argument and theory does not postulate the actual existence of possible worlds, nor does it defend a view of possible worlds. Their use of this phrase in this instance is synonymous with 'a different scenario', or, 'a different sequence of events'. 'Possible worlds' here is used for the sake of clarity in understanding.

own steering wheel which he will use to override Sally's and turn the car to the left should Sally try turning to the right. What Fischer and Ravizza (1998) demonstrate with this example is that these two types of control come apart. Even though Sally lacks the regulative control to do otherwise because of the existence of the driving instructor's potential intervention, she is still morally responsible for turning to the left because she still possesses guidance control. (Fischer & Ravizza, 1998).

To summarise thus far, Fischer and Ravizza's (1998) theory still maintains that an agent requires some degree of control to be morally responsible.[11] By positing distinct types of control, Fischer and Ravizza demonstrate that we can still be morally responsible even if it is the case that regulative control is incompatible with determinism. They argue that guidance control is the kind of control necessary for moral responsibility, and that this kind of control is compatible with determinism (Fischer & Ravizza, 1998). On a Fischer and Ravizza framework, if determinism is true we do not lose moral responsibility; we might just lose the ability to do otherwise. Such a successful philosophical manoeuver means that we can have a theory of moral responsibility which side steps one of the major arguments for theories of Incompatibilism. I will discuss the significance of this later on in the chapter.

I move now to show, as Fischer and Ravizza (1998) argue, that simply possessing guidance control is not enough to be granted moral responsibility; that it is threatened by 'manipulation arguments'. Having done so, I will proceed to examine and explain Fischer and Ravizza's posited solution of a reasons-responsive mechanism.

Suppose in 'Assassin' that Jack uses his device to make Sam kill the mayor (Fischer & Ravizza, 1998). Sam possesses guidance control, but it does not seem appropriate to say he is

[11] This is fortunately consistent with assumptions made by other moral theories and everyday intuitions. I will discuss the significance of this towards the end of the chapter.

morally responsible. It also does not seem appropriate to say that agents acting under the influence of drugs, hypnotism, direct brain manipulation, indoctrination, or irresistible psychological impulses are morally responsible.[12] Fischer and Ravizza (1998) posit that such an agent is not responsive to reason, and it is being reasons-responsive which makes us morally responsible.[13]

Being reasons-responsive means that if the agent were presented with sufficient reasons to act otherwise, then, at least in some cases, the agent would recognize these reasons and act otherwise (Fischer & Ravizza, 1998; McKenna, 2010). If an agent were to act as they do regardless of what reasons were presented to them then they are not reasons-responsive and therefore not morally responsible (McKenna & Coates, 2015). However, Fischer and Ravizza (1998) draw another distinction here.

Consider again, 'Assassin'. While Sam the *agent* is not reasons-responsive (because he would end up killing the mayor despite what reasons were presented to him), Sam's action is the product of a *deliberative mechanism* that *is* reasons-responsive[14] (Fischer & Ravizza, 1998). Fischer and Ravizza (1998) claim that it is specifically the agent's *deliberative mechanism* which must be reasons-responsive, and if it is then the *agent* may be held responsible for actions resulting from it. Therefore the clause changes; filtering out any counterfactual interveners, if

[12] Fischer and Ravizza (1998) do acknowledge that this is, of course, provided that the agent is not responsible for being in that state in the first place. This is what they refer to as a 'tracing' element (Fischer & Ravizza, 1998, p. 50). I discuss the implications and applications of the tracing element at length in chapter four.

[13] Fischer and Ravizza argue that 'when there is normal, unimpaired operation of the human deliberative mechanism, we suppose that the agent is responsive to reason' (Fischer & Ravizza, 1998, p. 37)

[14] By 'deliberative mechanism' I mean the narrow spectrum of reasons and processes which are involved in the causal production of action (McKenna & Coates, 2015). Alternatively, a 'decision making and action producing' mechanism.

sufficient reason to do otherwise was presented to the mechanism, it would recognize those reasons and do otherwise (McKenna & Coates, 2015).[15]

Before continuing on to discuss the notion of reasons-responsiveness in more detail, it is important that I iterate another criterion for moral responsibility mentioned by Fischer and Ravizza (1998) which also affects manipulation arguments; that is, the deliberative mechanism that the agent acts on must be the *agent's own*. According to Fischer and Ravizza, it is not merely enough to be acting from just any deliberative mechanism. Consider another scenario, where James is an agent who has also had a brain-manipulating chip implanted in his head by a scientist, and the scientist manipulates James to hit his neighbour with his car. In this scenario, Fischer and Ravizza argue that James cannot be responsible for this action because (among other factors) this action has not issued from his *own* deliberative mechanism, but from the mechanism of the chip. Thus it is a necessary requirement on Fischer and Ravizza's account that an agent act on their own reasons-responsive mechanism. This is the 'ownership' condition.

According to Fischer and Ravizza's (1998) ownership condition, an agent makes a deliberative mechanism his own by 'taking responsibility' for it. In their (Fischer & Ravizza, 1998) text, the author's specify three necessary and sufficient conditions which the agent must satisfy in order to take responsibility for their mechanism, thereby making it their own. Firstly, the subject must view themselves as an *agent*, and see that their actions in the world will have certain implications and upshots. Secondly, the agent must view themselves as an apt target for

[15] As I make clear later in this chapter though, it need not be the case that the mechanism would produce a different action in response to just *any* sufficient reason to do otherwise. A sufficient reason to do otherwise has its own set of requirements.

the 'reactive attitudes'[16] of others *given our actual social practices*[17]. Finally, it requires that the agent have a particular cluster of beliefs about oneself in light of the first two requirements set out here. That is, the agent must believe that it is at least *prima facie* plausible that he is an apt target for reactive attitudes, and he must be willing to put aside doubts for practical purposes.

These beliefs that the agent holds, say Fischer and Ravizza (1998), need not be explicitly voiced, nor need the agent be consciously aware of them. Instead, they argue that taking responsibility is a matter of having certain kinds of beliefs about oneself and one's actions. Thus, an agent makes a deliberative mechanism his own, and satisfies the ownership condition, by taking responsibility for that mechanism. This is achieved by having the right kinds of beliefs, namely, that they are an agent and their actions have implications in the world, and that they are an apt target for reactive attitudes in some circumstances given our social practices.[18]

Reasons-responsiveness consists of two criteria; receptivity and reactivity (Fischer & Ravizza, 1998). Receptivity is the ability to *moderately recognize* or acknowledge reasons (Fischer & Ravizza, 1998). [19] This entails recognizing a *moderately consistent or coherent pattern* of reasons (Fischer & Ravizza, 1998). 'Pattern' of reasons refers to how recognized reasons fit together and relate to each other; understanding that some reasons are better than others; and that accepting one reason as 'sufficient' implies we would also accept any stronger

[16] A reactive attitude is essentially those attitudes or emotional responses which others may feel in response to certain actions. Typically, (at least in the scenarios this project is describing), these reactive attitudes are directed at the agent who carried out said actions. In this case, to satisfy this requirement for the ownership condition, the agent must view themselves as an apt target for these reactive attitudes.

[17] Fischer and Ravizza explain that it helps to view the agent's actions as "a move in a social game". In other words, our society has a particular set of social standards and practices in place. Further, the agent's actions have upshots and implications. Therefore, the agent must be able to view themselves as an apt target for reactive attitudes depending on the way their actions interact with these social standards and practices.

[18] I return later in this project to discuss what these requirements mean for our implicit attitudes, and whether or not we can consider them, or the mechanism they feature in, the agent's own.

[19] Fischer and Ravizza distinguish between weak receptivity, strong receptivity, and moderate receptivity as degrees of consistency in the mechanism's ability to recognize reasons. In other words, 'weak' receptivity would entail poor consistency (recognizes few reasons, sporadically), and 'strong' receptivity would entail something akin to perfect consistency (or recognizing every reason).

(or better) reason as sufficient.[20] It must be noted that an agent is not required to be able to be receptive to *every* sufficient reason to do otherwise (Fischer & Ravizza, 1998); such would require perfect (strong) receptivity and is too stringent a requirement for human agents. However an agent must be able to be at least ‘minimally consistent’ (McKenna, 2010), and so sporadic or *weak* receptivity is too loose a requirement.[21] Thus a moderate requirement “gives rise to a minimally comprehensive pattern” of receptivity, while still not requiring us to recognize every reason to do otherwise (Fischer & Ravizza, 1998).

Reactivity is the ability to “translate reasons into choices” and then *act in response* to those recognized reasons (Fischer & Ravizza, 1998, p. 86).[22] In addition to the receptivity requirements, an agent is morally responsible if their *actual-sequence* mechanism[23] is capable of reacting to different reasons or incentives (Fischer & Ravizza, 1998). If the mechanism is properly reactive to reasons, then if a different (or sufficient) reason to do otherwise was presented, the mechanism could have responded to that different reason (Fischer & Ravizza, 1998). [24] The mechanism thus demonstrates it can react to different inputs, and so could have acted alternatively (McKenna, 2010).[25]

[20] Consider the following example: Suppose I want to attend a concert, but I can only do so if the tickets cost $100. In this case the tickets costing $100 is a sufficient reason for me to attend the concert. A ‘stronger reason’ would be if the tickets cost $90, or any amount less than $100. Proper receptivity requires I be able to recognize that, if $100 tickets is a sufficient reason to attend the concert, $90 would also be a sufficient reason to attend the concert. Thus my reasoning for purchasing a $90 ticket is *consistent* with my reasoning that $100 tickets is sufficient.

[21] If we specified weak receptivity, then the agent may recognise “a pattern of reasons so odd that it seems clear that the agent is not morally responsible” (Pereboom, 2006, p. 199).

[22] Consider the concert ticket example: If my deliberative mechanism is appropriately reactive then I can translate the fact that tickets are $90 in to the decision to buy the tickets, and then go to the concert. In this way I have translated a recognized, sufficient, reason for acting in to a course of action and then carried out that action.

[23] The notion of ‘actual-sequence’ also applies to the receptivity component. By ‘actual-sequence’ Fischer and Ravizza refer to the *sequence* of events that *actually* happen. In other words, the actual-sequence mechanism is the deliberative mechanism that is actually in operation in the real world for the agent. It is not the hypothetical sequences of possible alternative events which we would postulate take place in the ‘other worlds’. Thus the actual-sequence is what actually happens.

[24] Fischer and Ravizza require that if the mechanism is going to act otherwise then it must be *because* of the sufficient reason to do otherwise. That is, the reason to do otherwise and the doing otherwise must be “appropriately connected” (Fischer & Ravizza, 1998, p. 64)

[25] Such as by not performing a morally negative action

Fischer and Ravizza (1998) also assert the claim that “reactivity is all of a piece”. That is, the mechanism is not required to react to *every* reason to do otherwise in order to demonstrate eligibility for moral responsibility. They believe if the mechanism can react to one reason to do otherwise then this demonstrates that the mechanism can react to *any* reason to do otherwise (Fischer & Ravizza, 1998). [26] In other words, if the mechanism can react to *at least one* sufficient reason to do otherwise, then it demonstrates that, in the world in which the agent does wrong, it possesses “the general capacity to react to reasons” (McKenna, 2010).[27] In this regard, Fischer and Ravizza posit a *weak reactivity* requirement[28]. Therefore, the agent “can still be morally responsible if they are receptive to sufficient reasons to do otherwise and choose not to react to those reasons” (Pereboom, 2006, p. 199).

Fischer and Ravizza (1998, p. 77) also specify that while the mechanism must be receptive to an appropriate range of reasons, at least some of these reasons must be *moral* reasons.[29] If the mechanism is not able to appreciate, understand, or be receptive to moral reasons then “agents would fail to be *moral* agents at all” (Fischer & Ravizza, 1998, p. 77). Again, the agent need not recognize every moral reason; merely a consistent and coherent pattern, and recognize that they *apply to the agent* (Fischer & Ravizza, 1998). It is for this reason that animals, psychopaths, and young children[30] are not morally responsible agents on the Fischer and

[26] Italics not in original text.

[27] In other words, by being able to react to a sufficient reason to do otherwise, the mechanism demonstrates that it possesses this ‘general capacity’; by reacting to one it shows it can react to any.

[28] As it was with receptivity, so too is weak reactivity the requirement that the mechanism be able to react to at least one reason to do otherwise. As Fischer and Ravizza argue at this point, weak reactivity is sufficient for moral responsibility. Strong reactivity would require the agent to react to almost every reason to do otherwise, and again this is too stringent a requirement; asking for perfect responding.

[29]A ‘moral’ reason is one which reflects an understanding and appreciation of the rights or interests of others, and that they may be greater or stronger than our self-interests (or, ‘prudential’ reasons).

[30] Depending on their stage of development

Ravizza account. While these agents act on incentives, and are receptive to reasons, they are not able to consistently recognize or react to moral reasons (Fischer & Ravizza, 1998).[31]

Fischer and Ravizza's (1998) theory thus far *seems* to place an emphasis on the utilisation of practical reasoning[32]; the agent must recognise reasons, evaluate reasons, translate them in to action, and then carry them out. But not all actions are the product of practical reasoning; some actions operate outside of conscious awareness, such as habits, customs, or instincts (Fischer & Ravizza, 1998). However, "nowhere…is it required that the mechanism use 'practical reasoning" (Fischer & Ravizza, 1998, p. 85). Even without operating consciously, the deliberative mechanism can still be sensitive to, and react to, reasons.

Suppose you drive the same route to work every day and eventually you start taking the same exit out of habit, even though there is another turn-off further down the road.[33] One day your habitual turn-off is closed. Without thinking about it you continue along the road and take the second exit. In this example the mechanism has not utilised practical reasoning (at least in the sense that the agent is conscious of, or deliberately carrying out, the deliberative process). Nonetheless the mechanism has recognised reasons for acting otherwise than normal, and has reacted to them. Therefore, you are still responsible for taking the second exit. Fischer and Ravizza's (1998) account still allows us to ascribe responsibility for 'non-reflective behaviour'.

To summarise, merely possessing guidance control is not enough to be held morally responsible. For instance, an agent acting under the influence of drugs, direct brain manipulation, hypnosis, or irresistible psychological impulses or desires, for instance, may be the one carrying

[31] In some cases, psychopaths may be able to understand that moral reasons, or reasons stemming from the interests and rights of others, exist. However, they are not necessarily able to understand that these reasons apply to them. So they fail to be moral agents.

[32] While there may be differing opinions on what constitutes 'practical' reasoning, here it is used to suggest conscious reasoning

[33] This example is taken from Fischer & Ravizza (1998). This example is their 'driving down University Avenue' example.

out the action, however we might say he *lacks* guidance control and may not be morally responsible. Fischer and Ravizza (1998) specify this is because their actions do not issue from a moderately reasons-responsive mechanism.[34] This mechanism, operating in the actual-sequence, must be receptive to a consistent pattern of reasons to do otherwise (including at least some moral reasons). It must also be capable of reacting to at least one of these sufficient reasons because of that sufficient reason. An agent is morally responsible if, via a reasons-responsive mechanism, *chooses* to act (or chooses to not act) on the basis of a sufficient reason to do otherwise (McKenna, 2010, p. 47).

Having sufficiently detailed and examined Fischer and Ravizza's (1998) theory, I move now to consider some queries and questions raised against it, and determine whether Fischer and Ravizza respond appropriately. From the above, we observe Fischer and Ravizza arguing for what McKenna (2010) refers to as a 'striking asymmetry thesis'; while moral responsibility requires the mechanism be moderately receptive to reasons, it need only react to at least one reason to do otherwise. While this asymmetry is indeed crucial to the success of their account, McKenna argues that it need not be so extreme; that such an *extreme* asymmetry invites "controversies on several fronts" (Pereboom, 2006, p. 200).

McKenna (2010) poses the following example. Suppose Michelle works on her garden from a mechanism that is moderately receptive to reasons. This would entail that she would recognise Michael's losing a finger, a foot, or having a heart attack, as sufficient reason to do otherwise and not work in the garden. Fischer and Ravizza (1998) require only that she *react* to but one of these reasons; as long as she recognises that the other scenarios would be sufficient

[34] Specifically, Fischer and Ravizza (1998) explain that drugs, direct brain manipulation, hypnosis, or genuinely irresistible psychological impulses or desires give rise to a *new* deliberative mechanism which is not responsive to reason. In some cases, when the agent is operating on a new deliberative mechanism, we might say he is not morally responsible because his degree of guidance control has been undermined by the aforementioned influences. Cases such as these are not limited to the few examples I have listed above.

reasons to act otherwise as well, “it need not be the case that in ‘other worlds’ she acts upon those reasons” (McKenna, 2010, p. 98).

McKenna (2010) argues that the idea that Michelle would react to only one of these reasons, and no *similar* reasons (as Fischer and Ravizza (1998) seem to suggest), implies that Michelle is not a proper moral agent. This is because, if she truly appreciated the moral import of these other reasons, then she ought to react differently “in at least *some* other similar worlds” where these reasons are present (McKenna, 2010, p. 98)[35]. It is true that for Fischer and Ravizza’s theory to function as they need it to that they make allowances for the agent to recognise sufficient reasons and not act[36]. So while they need this asymmetry, McKenna suggests that Fischer and Ravizza should make the amendment that moral responsibility requires moderate receptivity and *weaker* reactivity[37]. As I believe this to be sound advice, I believe Fischer and Ravizza have responded appropriately and rightly to this suggestion, as McKenna recognises in footnote 23 that, through personal correspondence, Fischer has at least acknowledged this point.

Pereboom (2006) (and Mele, 2000 to an extent) voice a criticism along similar lines to McKenna (2010), pertaining to the claim that ‘reactivity is all of a piece’. Mele posits the following example. Fred is an agoraphobic. Because of his crippling fear, he decides not to leave his house and misses his daughter’s wedding. However, if Fred’s house were on fire, then Fred would have recognised this as a sufficient reason to leave his house, and would have attended the wedding. On Fischer and Ravizza’s (1998) account, because there is *some* situation in which Fred would do otherwise, Fred’s deliberative mechanism meets the weak-reactivity criteria, and Fred is morally responsible for missing his daughter’s wedding. Mele and Pereboom disagree. If his

[35] Remember that Fischer and Ravizza specify that a moral agent must be able to recognise, appreciate, and understand reasons to do otherwise that have significance for others, and not just self-interest.

[36] This is because it allows them to ascribe moral responsibility to those agents who *knowingly* do wrong; who knowingly make the wrong choice.

[37] ‘Weaker’ as opposed to ‘weak’ reactivity. This is slightly more consistent responding than can be achieved if we simply require weak reactivity. This way the agent will respond to *similar* reasons to do otherwise, not just one.

agoraphobia is so crippling that the only way to make him leave is through something as drastic as a raging house fire, then it does not seem appropriate to say he is morally responsible.

Pereboom (2006, p. 200) argues that it is "unintuitive to think that the mechanism by which he decides to stay home can react to *any* reason to do otherwise" just because it can react to the one reason presented by the fire. Therefore, how come a mechanism can respond to a reason to do otherwise in one situation but not others? Fischer and Ravizza (as cited in Pereboom, 2006) claim this is because in some situations the agent gains more energy and focus from some reasons than others (Pereboom, 2006). This acquisition of energy or focus gives rise to a different mechanism, and on this new mechanism it will react to those reasons (Pereboom, 2006). Therefore Fischer and Ravizza amend that, *holding fixed the actual mechanism*, reactivity is all of a piece (Pereboom, 2006). [38]

However, Fischer and Ravizza's (as cited in Pereboom, 2006) response seems inadequate and unintuitive. Pereboom (2006) and McKenna (2001, as cited in Pereboom, 2006) rebut that it does not seem implausible to think that the mechanism would remain the same regardless of changes in energy or focus. It seems like Fischer and Ravizza's individuation of mechanisms is a response driven by the needs of the theory, and not by the actual nature of the mechanism[39] (McKenna, 2001, as cited in Pereboom, 2006). So should Fred be held morally responsible for missing the wedding? McKenna asserts that there is no good answer; Fischer and Ravizza's (1998) theory has to ascribe moral responsibility.[40]

[38] This response of Fischer and Ravizza is described in Pereboom (2006)

[39] In other words, Fischer and Ravizza's response sounds more like a claim about what the mechanism *would need to do* in this situation to avoid Pereboom and Mele's challenge, as opposed to what it *actually* does or how it *actually* operates.

[40] While some may argue that this response from Fischer and Ravizza is unfair, I argue that even if they 'bite the bullet' here, and assert that Fred is morally responsible, that even then the consequences may not be too severe. In this scenario, Fred might be said to be morally responsible for missing his daughter's wedding, given how he meets the criteria for reasons-responsiveness set out by Fischer and Ravizza. However, I believe Fischer and Ravizza would still assert that Fred was not *blameworthy* for having missed the wedding, and by maintaining this distinction, would make this assessment of Fred seem much more reasonable. After all, Fred meets the requirements set out

A challenge that may be raised at this point would be just how much responsiveness does the mechanism have to display in order to satisfy the conditions for moral responsibility? I argue here that it would be a frivolous and impossible task to establish a principled response to this question. Of course, a certain degree of responsiveness is required to satisfy the conditions for moral responsibility.[41] But again, this is a matter of degrees; blameworthiness and moral responsibility all are discussed in terms of degrees.[42] Therefore, as the degree to which the deliberative mechanism is responsive to reason increases, so too does the degree to which the agent can be said to be morally responsible. Likewise, as a mechanism becomes less responsive, the degree to which the agent can be held morally responsible declines. Consequently, it seems more appropriate to do as Fischer and Ravizza have done (and to an extent, as Pereboom (2006) and McKenna (2001) have argued for reactivity) and provide a more general heuristic to evaluate the degree to which a person meets the reasons-responsiveness criteria, as opposed to attempting to specify a 'hard-line' threshold requirement for responsiveness.[43]

What comes out of our inquiry into the Fischer and Ravizza (1998) theory is an understanding that the ability to do otherwise is not what makes an agent morally responsible. Instead, an agent is morally responsible if she acts from her own, moderately reasons-responsive deliberative mechanism. This requires the agent's deliberative mechanism to be receptive to a

by Fischer and Ravizza for moral responsibility, but it does not seem appropriate to blame Fred for being *unable* to react to a pattern of reasons in light of his psychological circumstances.
Fred meets the requirements set out for reasons-responsiveness; while Fred is only reacting in response to one reason, Fred none the less is receptive to, and recognises, a consistent pattern of reasons

[41] This responsiveness must still adhere to a consistent and coherent pattern of reasons

[42] Hence why we may say someone is more or less blameworthy for an act than someone else; or is more or less morally responsible for an act than someone else

[43] While it might be argued that a response like this, of postulating a heuristic as opposed to a threshold merely pushes back the initial question of 'how much responsiveness is enough', it is not an uncommon or unique response. We observe this kind of vagueness across other areas of philosophy as well, including other accounts of moral responsibility, epistemology, bioethics, etc. That being said, there are of course going to be instances when an agent's mechanism is not going to be responsive enough to satisfy moral responsibility requirements. Despite the advocacy of a general heuristic, I believe that our intuitions ought to suffice in informing us of when the degree of responsiveness is too low to allow an agent to be morally responsible.

moderate range of reasons (some of which are moral reasons) which fit together in a coherent pattern. It also requires that the deliberative mechanism be able to react or respond to a weaker array of reasons; the mechanism need not respond to every reason that it recognises but it must be able to respond to some, not just one. The recognising and reacting to reasons also need not be occurring in conscious awareness, and so we can be held morally responsible for non-reflective behaviours on this account, but it must be occurring in the actual-sequence (not in hypothetical, alternative sequences of events).

Upon close inspection, it is clear that there are numerous benefits and strengths to the Fischer and Ravizza (1998) account as a theory of moral responsibility. Firstly, as I mentioned early on in this chapter, by drawing a distinction between regulative and guidance control, and demonstrating that guidance control is both compatible with determinism and required for responsibility, Fischer and Ravizza's theory attempts to work around a major argument within the free will debate. As moral responsibility has been thought in the past to require the ability to do otherwise[44], discussion of moral responsibility can sometimes be precluded by a major position in the free will debate. Incidentally, we can sometimes struggle to talk about moral responsibility because we cannot get past free will. Fischer and Ravizza show with this theory that we can talk about moral responsibility without talking about free will. Furthermore, as this theory is compatible with determinism, Fischer and Ravizza provide the determinist with moral responsibility; a sometimes difficult accomplishment.

Another strength of the Fischer and Ravizza (1998) theory is its realistic requirements. This theory is very much linked to the psychological capabilities of human moral agents, and requires only faculties which can be reasonably expected to be possessed by most agents, such as sound operation of the mind, and the ability to recognise reasons. Moreover, this theory also

[44] A distinctively incompatibilist notion, and an often presumed characteristic of free will

appreciates that people's ability to be receptive and respond to reasons consistently is not perfect. As such it does not place idealistic or unobtainable requirements on the moral agent. By drawing the asymmetry as McKenna (2010) describes it, Fischer and Ravizza make allowances for humans to behave imperfectly, but still be held morally responsible where it matters. Thus the theory seems to fit not only with our intuitions about responsibility, but also our current understanding of psychology.

As well as being realistic, I believe this theory to be quite practical as well. The theory does not require being able to do otherwise. Instead it grounds moral responsibility in the actual-sequence; what the agent *actually* does, as opposed to being based on what other options were available to the agent. This means that we can focus on what the agent actually did and whether they had control in that instance. In a real life setting I believe this is important because it means that ascribing responsibility can focus on the reality of the situation and not be hindered by hypothetical alternative scenarios; of particular importance in criminal prosecutions. Finally, this theory also functions with theories of punishment as conditioning exercises, because the punishment can be structured so as to revise how agents recognise, appreciate, and respond to reasons.

What I hope to have achieved thus far is an in-depth examination into the motives, arguments, and criticisms of the Fischer and Ravizza (1998) theory of moral responsibility. I hope to have shown that moral responsibility does not require the ability to do otherwise. Instead, being morally responsible requires operating from a moderately reasons-responsive deliberative mechanism. In the following chapters I will be applying Fischer and Ravizza's theory to the challenges presented by implicit attitudes, and asking whether we can be held morally responsible for our actions which issue from, or are caused by, our implicit attitudes.

Chapter Two – Implicit Attitudes

In chapter one I provided an inquiry and examination in to the Fischer and Ravizza (1998) account of moral responsibility. I demonstrated that by focusing on a reasons-responsive deliberative mechanism, Fischer and Ravizza manage to provide an account of moral responsibility that is both realistic in its requirements, and practical in its applications. It was established that moral responsibility requires not merely guidance control over one’s actions, but also that those actions issue from a deliberative mechanism which is capable of recognizing a coherent and consistent pattern of reasons (including at least some moral reasons), and which is capable of responding to at least some of those reasons. If the mechanism can do this, then we may hold the agent morally responsible.

For the second component of my project, I seek to apply the Fischer and Ravizza (1998) theory of moral responsibility to the psychological construct, ‘implicit attitudes’. Specifically, my project concerns the interaction of implicit attitudes with behaviour on a Fischer and Ravizza theory of moral responsibility. As I will proceed to argue in the proceeding chapters, implicit attitudes can play a role in the deliberative mechanism issuing in our behaviour, and can therefore affect the actions we perform. However, the exact nature of implicit attitudes, their operation, and the extent to which they affect behaviour (if at all), is a matter of great debate in the empirical and philosophical literature. Opinions are varied and conflicting. Therefore, in this chapter, I aim to shed a clarifying light on this topic.

What I attempt to provide here is a brief overview of the literature surrounding implicit attitudes. This review is by no means intended to solve the challenges and questions raised by implicit attitudes. Rather, I intend to highlight what I regard to be some of the more pressing ideas regarding implicit attitudes.[45] Thus the purposes of this chapter will be, firstly, to introduce implicit attitudes, and examine their nature and operation, and how they differ from explicit attitudes. Secondly, I will examine two of the foremost testing procedures used to examine implicit attitudes, and discuss the motivations behind developing implicit and explicit measures. Finally, in light of the evidence provided by implicit and explicit measures, I will introduce some pieces of research which argue whether we should consider implicit and explicit attitudes as dual-constructs, or as a single construct that is measured in different ways.

As I have already mentioned, the nature of implicit attitudes is a matter of contention. Attitudes more broadly are characterised by the fact that they provide us with a valanced (or, a favourable or unfavourable) summary and evaluation of aspects of our environment (Fazio & Olson, 2003; Wilson, Lindsey & Schooler, 2000). Our attitudes serve a functional role in that they predispose us to certain types of responding and behaviour when we encounter 'attitude objects' (Wilson, Lindsey & Schooler, 2000)[46]. My behaviour towards people, for instance, will be shaped by my attitudes towards them; if I like the person, if I hold a positive attitude towards them, then I may be disposed to treating them nicely.

A number of theories posit that attitudes are open to introspection; that they are readily accessible by the agent[47], that they are reportable, and that they can be subject to reasoning (Fazio & Olson, 2003; Greenwald & Banaji, 1998; Karpinski & Hilton, 2001; Wilson, Lindsey & Schooler, 2000). Attitudes fitting this criterion of being accessible and reportable are often referred to, however, as 'explicit attitudes', and are the ones we are usually consciously aware of. Rydell, McConnell, Mackie and Strain (2006) characterise them as attitudes which are both

[45] This does not mean that these are the *only* pressing ideas. The literature surrounding implicit attitudes is broad and relatively new. There are many questions and challenges to be addressed and I cannot hope to address each one of them in this project.

[46] 'Attitude objects' are things in the world which we may hold an attitude towards, including, but not limited to, people, places, jobs, etc.

[47] Specifically, 'consciously' accessible

reportable and consciously controllable. Because of this, explicit attitudes have been examined using very 'direct' testing procedures. While these testing procedures are "tried and tested" (Cunningham, Preacher & Banaji, 2001), these methods make a number of potentially problematic assumptions.

Acquiring an accurate measure of an agent's explicit attitudes proves difficult, as explicit attitudes are affected by a number of 'artifacts'[48] which can confound results (Greenwald, et al., 2002). A subject being asked to report their explicit attitudes seems affected by such influencing factors as impression management, demand characteristics, and social desirability (Dovidio, Kawakami, Johnson, Johnson & Howard, 1997; Greenwald et al., 2002; Karpinski & Hilton, 2001). The presence of factors such as these in a testing environment can dispose subjects to change their responses in an effort to please the examiner, or preserve their self-image or social standing, by not reporting their real, or potentially socially-undesirable, attitude.

Incidentally, these procedures also necessarily assume (and require) that the agent is capable of reporting honestly and does so (Cunningham, Preacher & Banaji, 2001; Greenwald et al., 2002). Because these measurement instruments (such as self-report) can be affected so significantly by these artifacts, and because getting consistent results from within subjects when these variables are changed is so problematic, researchers developed alternative, 'indirect' testing procedures, designed to measure explicit attitudes but be unaffected by the aforementioned artefacts.

These indirect methods have taken many forms, including (but not limited to) facial electromyography, fMRI, eye blink response, cardiovascular reactivity, and others (Fazio & Olson, 2003). Foremost among these procedures has been the emergence of the Implicit

[48] By 'artifacts', Greenwald et al. (2002) refer to factors which confound and affect accurate responding from the subject. I go on to list some examples of confounding artifacts.

Association Test (IAT) developed by Greenwald & Banaji (1995), and such priming methodologies as the Affect Misattribution Procedure (AMP) developed by Payne, Cheng, Govorun & Stewart (2005). The general idea behind these approaches is that they are able to circumvent social desirability concerns and self-representation ideals by not requiring direct verbal or written report from the subject (Cunningham, Preacher & Banaji, 2001; Fazio & Olson, 2003). Moreover, some associative links which self-reports measure may simply not be available to introspection (Greenwald, et al., 2002). The principle of these procedures is that, by not directly asking the subject for a response, the subject is presumably not able to exercise direct control over influencing their responses because they are not aware of what is being tested (Fazio & Olson, 2003). [49] These testing procedures rely on quick, indirect, automatic responding rather than direct and controlled responding in order to more accurately gauge the subject's 'honest' attitudes.

In using both direct and indirect procedures for measuring attitudes, studies have found the subject to respond differently. For instance, subjects may respond in one way to a question when asked on a direct measure, and yet may provide a completely different (even contradictory) response if asked on an indirect measure. I will return to discuss this matter later in the chapter. However, the disparity in the results provided by direct and indirect measures have led to the positing of the dual existence of automatic, 'implicit' attitudes alongside our explicit attitudes.[50] These results seem to indicate that some attitudes can exist outside of conscious awareness and control.

[49] This is not to say that subjects are never (or cannot become) aware of what is actually being measured by these procedures. In a number of cases subjects may actually become aware that the content of their attitudes are being examined. The idea behind these procedures is that they measure more automatic responses which reduces the opportunity for subjects to exert deliberate control, and potentially bias their responses.

[50] As I discuss at towards the latter half of this chapter, there are many possible interpretations of what the disparity between these testing procedure really means. This is merely one common interpretation.

Furthermore, unlike explicit attitudes, the content of implicit attitudes may be introspectively inaccessible to the agent (Karpinski & Hilton, 2001).[51] One of the original descriptions of what implicit attitudes are comes from the work of Greenwald, McGhee & Schwartz (1998) who identified them as being 'introspectively unidentified (or inaccurately identified) traces of past experiences'. Rydell, McConnell, Mackie & Strain (2006) propose a similar understanding, going so far as to suppose they are "attitudes to which people do not initially have conscious access and whose action cannot be controlled".

Emerging from the literature is a seemingly consistent theme; the manner in which we are aware of our implicit attitudes is different to how we are aware of our explicit attitudes. In some ways it seems to be the case that we are quite consciously aware of our explicit attitudes, but we are at least originally less aware of our implicit attitudes and their content. Nonetheless, what one may gleam from an inspection of the literature is that there exists at least an intuitive distinction between implicit and explicit attitudes. Whether or not such a distinction is justified is a matter I will address shortly. For now, a few characteristics of implicit attitudes seem typically agreed upon; implicit attitudes are a type of mental state which possesses evaluative content in regards to attitude objects. Furthermore, implicit attitudes are automatically activated; they can come to mind with little conscious effort or intention (Fazio & Olson, 2003; Payne, Cheng, Govorun & Stewart, 2005; Rydell, McConnell, Mackie & Strain, 2006; Perugini, 2005; Wilson, Lindsey & Schooler, 2000).

Other characteristics of implicit attitudes may not be so widely agreed upon. Studies conducted by Payne, Cheng, Govorun & Stewart (2005) indicate that inhibiting the expression or influence of implicit attitudes may be difficult. At the same time, other studies have said that it is

[51] The degree to which implicit attitudes are introspectively inaccessible is another contentious issue in the literature, as studies have suggested that subjects might be aware of the contents of their implicit attitudes after all, or may come to be aware of their content with the right instruction (Payne, Cheng, Govorun & Stewart, 2005). Even so, it remains that implicit attitudes are at least initially less introspectively inaccessible than explicit attitudes.

possible to inhibit the manifestation of implicit attitudes in behaviour, and that there are a number of ways one can do so (Holroyd, 2012). Overall, it is clear to see that, exactly what implicit attitudes are, is hard to say. However, I will return to this shortly.

As I have remarked previously, the purpose of this study is to examine the interaction, and any possible challenges that might arise, between implicit attitudes and moral responsibility. What brings implicit attitudes in to the domain of moral responsibility is that they can affect behaviour. That is, our actions are in some part moderated by our implicit attitudes (Dovidio, Kawakami, Johnson, Johnson & Howard, 1997; Karpinski & Hilton, 2001; Levy, 2014; Perugini, 2005)[52]. Given there are still conflicting opinions on this topic, it seems that we must at least consider the possibility that implicit attitudes contribute to the deliberative mechanism that issues in behaviour, even if the resulting behaviour does not reflect the content of said implicit attitude.[53]

Again, what makes implicit attitudes a moral concern is that they affect behaviour. Given that we can hold negatively valanced implicit attitudes regarding gender, religion, race, and more, implicit attitudes can make a significant contribution to the expression of a variety of harmful and derogatory behaviours, and in a number of environments. Discriminatory implicit attitudes can be expressed in employment settings (affected hiring rates and performance evaluations), education (disciplinary actions, evaluations of tests), health care, (treatment in triage, prescription administration), and many others (Greenwald, Banaji & Nosek, 2015). These negative implicit

[52] This is yet another characteristic of implicit attitudes which is debated among philosophers and psychologists. While some studies claim to have found little to no influence of implicit attitudes on behaviour, there are a number of other studies which implicate implicit attitudes as not only a contributing factor to the mechanism issuing in behaviour, but that they may sometimes be the main guiding force.

[53] For example, I may hold the implicit attitude that black people are inferior to white people. This implicit attitude may affect my behaviour; it may influence me or dispose me to be rude and inconsiderate in my relations with black people. However, due to the contribution of other factors in the mechanism issuing in behaviour, I may not behave this way; I may be polite and well-spoken towards black people. In this sense my implicit attitude has played some sort of role in the production of my behaviour, even though my behaviour does not reflect my implicit attitude.

attitudes may result in unfair treatment of individuals in real life settings, and this is particularly problematic when they apply to many people, or when it applies repeatedly to the same person (Greenwald, Banaji, Nosek, 2015; Levy, 2014).

Therefore, when an agent performs discriminatory or abusive behaviour toward another agent, one question we might feel compelled to ask is whether that person is influenced by the implicit attitude when they act or not. Our moral theory requires that agents have a particular type of control over their actions, and that their deliberative mechanism be able to recognise and consider reasons for acting. Factors such as the nature of implicit attitudes, and the extent of the contribution they make to the deliberative mechanism issuing in the discriminatory behaviour, will affect how responsible we may be for their manifestation in behaviour.

At this point I would like to take a brief moment to summarise what I have presented thus far. Attitudes are valanced summaries of, and responses to, attitude objects, and can play some sort of role in the deliberative mechanism producing our behaviour. Our attitudes can be broken down in to two categories. Explicit attitudes are largely deliberative attitudes; they are available to introspection; we are consciously aware of them; and they are typically thought to contribute to controllable behaviours. Then there are implicit attitudes which are more controversial. Agents are typically to some extent unaware of their implicit attitudes and their contents; they seem to operate outside of conscious awareness, but do so automatically; they too can be found to impact (predominantly automatic) behaviour. Because implicit attitudes can contribute to the deliberative mechanism which issues in specifically discriminatory behaviour, they are made a moral concern, and merit our attention as moral theorists.

Moving forward, I will examine two of the foremost testing procedures used to analyse these implicit attitudes; namely the Implicit Association Test (IAT) and the Affect Misattribution Procedure (AMP). Probably the most well-known, and arguably the most reliable test for measuring implicit attitudes is the IAT (Perugini, 2005; Rydell, McConnell, Mackie & Strain, 2006; Cunningham, Preacher & Banaji, 2001). The IAT was developed as, and intended to be, a method of indirectly measuring the strength of underlying associative structures between concepts; particularly, how associated an attitude object is with a certain attitude (Greenwald & Banaji, 1995; Greenwald, McGhee & Schwartz, 1998; Perugini, 2005).[54]

To do this, the IAT compares the length of time taken to associate positively-valanced words (e.g. 'beautiful') and negatively-valanced words (e.g. 'ugly', 'poison') with different stimuli based on what domain is being tested (for instance, for race, the stimuli might consist of black faces and white faces). Subjects are asked to pair the positively-valanced words with the white faces and the negatively-valanced words with the black faces. The stimuli are counter-balanced on subsequent trials such that positively-valanced words are paired with black faces and negatively-valanced words are paired with white faces. How quickly a subject matches these different combinations is supposed to reveal their implicit attitudes.[55]

Another commonly used implicit measure is the AMP constructed by Payne, Cheng, Govorun, and Stewart (2005); known as a priming procedure[56]. This procedure operates on the idea that agents typically imbue ambiguous stimuli with personal sources of meaning (Payne, Cheng, Govorun & Stewart, 2005). These 'personal sources of meaning' are taken in the AMP to reflect the subjects' implicit attitudes.

In this procedure, people are presented with a priming stimulus, like a picture, which elicits an initially positive or negative judgement. Subjects are then asked to evaluate a 'judgment

[54] While some may argue exactly what results of the IAT actually reveal about mental content, this is the way it was originally intended to be interpreted

[55] Say if I was quicker at matching negative words with a black face compared to matching negative words with a white face. Supposedly, this indicates that I hold a more negative implicit attitude towards black people than I do towards whites because I more readily associate negative words with black faces than I do white faces.

[56] 'Priming' is a procedure used to make certain actions more likely to occur by presenting the subject with a particular stimulus

target' (another picture), which is emotionally ambiguous, and are asked to avoid the influence of the prime in their evaluations. Depending on the evaluation of the prime, we ought to expect a similar evaluation to have occurred towards the ambiguous attitude object (Payne, Cheng, Govorun & Stewart, 2005).[57] In this procedure, misattribution ought to only occur if the subject is unable to monitor and control the influence of their attitudes for the prime in their judgements of the target. Therefore, misattribution is taken to likely reflect an "automatic or implicitly held" attitude. This is because the agent cannot control or inhibit the expression of their implicit attitudes.[58]

An examination of the procedures themselves reveals that these implicit measures have high reliability scores, correlate highly with each other (Cunningham, Preacher & Banaji, 2001) and, after accounting for measurement error, are found to be consistent across time and measures (Greenwald, McGhee & Schwartz, 1998). The IAT also demonstrates a near absence of moderating effects (Greenwald, McGhee & Schwartz, 1998). What is interesting though is that, while these procedures correlate highly with each other, the IAT and the AMP (as well as other implicit measures) correlate very poorly with explicit measures (Cunningham, Preacher & Banaji, 2001; Wilson, Lindsey & Schooler, 2000; Karpinski & Hilton, 2001; Fazio & Olson, 2003). When these implicit measures are performed alongside explicit measures, regarding topics of prejudice, stereotypes, smoking, food, alcohol, etc., we see a divergence emerge. Despite being asked similar questions, the responses subjects provide on implicit and explicit measures are different.

[57] This is made possible because of the human tendency to misattribute emotional cues. In this case, the agent will be disposed (or, 'primed') to evaluate the ambiguous judgment target in a particular way because of the way he evaluated the priming stimulus.

[58] Again, I wish to note that, while some may dispute that this is what results of the AMP demonstrate, this is the way it was originally intended to be interpreted

The second section of this chapter aimed to provide an examination of two of the foremost testing procedures used to examine implicit attitudes. I have presented a brief overview of the Implicit Association Test, and shown how the IAT might be used to examine the strength of underlying associative connections between implicit attitudes and associative objects. I then proceeded to examine the Affect Misattribution Procedure, and show how it utilises the human tendency to misattribute emotional cues to examine implicit attitudes. However, it was found that results of implicit measures and explicit measures, in response to similar questions, produce divergent and differing responses. That is, that there is a dissociation between the results of these two types of testing procedures; both of which are intended to measure an agent's attitudes. In the final section of this chapter, I provide a brief introduction to the arguments that have been raised in an effort to account for this discrepancy between test results.

There are a number of different ways in which this dissociation may be interpreted and a number of different theories have been put forward in response to these findings. Upon examining their original results from the IAT, Greenwald and Banaji (1995) interpreted this dissociation as suggesting that there is only one attitudinal construct, and the IAT and explicit measures tap different aspects of the one attitude. This is indicative of what Perugini (2005) refers to as the 'Independence of Measure' model. On this line of thought, an additive effect is observed, whereby both implicit and explicit measures "provide a unique contribution to behaviour prediction" (Perugini, 2005). As a result, depending on the context, and other contributing factors, the predictive power balance may change. In this case, maybe the implicit measure is the better predictor of behaviour; maybe in certain situations, the explicit measure is the better predictor of behaviour.

An alternative model argued for extensively by Wilson, Lindsey & Schooler (2000) is for the existence of 'dual attitudes'. They argue that the ambiguous dissociation between implicit and explicit measures indicates two different constructs; an implicit attitude and a simultaneously held, but distinct, explicit attitude. These distinct attitudes function differently and are responsible for different types of behaviour. For one thing, implicit attitudes are responsible for automatic and spontaneous behaviours, while explicit attitudes are more responsible for considered and deliberative behaviours (Perugini, 2005; Wilson, Lindsey & Schooler, 2000).

On a dual attitudes model, the difference between scores can also be attributable to the confounding artifacts that I mentioned earlier in the chapter. When asked about socially sensitive matters, such as race, religion, or gender biases, subjects may explicitly report a more socially acceptable attitude, even though they may actually hold a more prejudicial implicit attitude (Fazio & Olson, 2003). In these cases, the explicit attitude seems to be constructed on the spot as a way to 'save face' and preserve the subject's self-image. That is, on some level, it could be the case that subjects may be aware their implicit attitudes are socially undesirable, and may construct an explicit attitude which is more acceptable.[59]

The question that comes up when examining this possibility is, 'which is the real attitude?' (Fazio & Olson, 2003; Perugini, 2005). If it is the case that there are two attitudinal constructs, then which of these two measures is indicative of the 'real' attitude; the implicit or the explicit? This discussion ought to prompt two questions; (1) which is the real attitude, and (2) which attitude best predicts behaviour? A significant portion of the literature surrounding implicit attitudes seems to regard the second question as being the more pressing to ask. However, at the same time, this literature often takes the attitude which best predicts behaviour to be indicative of the agent's real attitude.[60] What research has indicated is that these two attitudes have an

[59] It is not always the case that agents are aware that they explicitly express an explicit attitude that is different to their implicit attitude. Given how implicit attitudes can exist and operate outside of conscious awareness, when an agent responds to a socially sensitive question in a socially desirable way, despite the existence of his implicit attitude, the agent may be unaware that he is behaving this way.

[60] I wish to express that I find the apparent 'conflating' of these two questions in this way to be odd, and that the reader should be alert to this oddity. Whether or not it is appropriate to think about implicit attitudes like this is a

'interactive effect' on behaviour; either one of them can be the guiding force on behaviour, but it depends on motivation of the subject.

If a subject has significant motivation to inhibit their implicit attitudes and express their explicit attitudes, then that is the attitude which will manifest in behaviour. In this case, while the term 'real attitude' is an ambiguous one, Fazio & Olson (2003) argue that it could refer to the degree to which that attitude predicts behaviour. Though this definition may be disputed, what this suggests is that, whether it is implicit or explicit attitudes that predicts behaviour will change in each situation depending on motivation.

In fact, the more socially sensitive the examined domain is, the greater the likelihood that these motivational factors will play some sort of role in responding (Fazio & Olson, 2003; Greenwald & Banaji, 1995; Greenwald, McGhee & Schwartz, 1998). Therefore, on the dual-attitudes model, the divergence between implicit and explicit attitudes is attributable to the motivation and opportunity the agent has to deliberate and explicitly override their implicit response (Fazio & Olson, 2003; Perugini, 2005). However, evidence for this view is inconclusive. Most of the time it appears we are simply *inferring* the existence of dual attitudes from our test results rather than testing for it directly (Perugini, 2005).

Thus, what comes out of an examination of this literature is a lack of a solid scientific conclusion supporting either the independence or dual attitudes model. Unfortunately, this dissociation is currently too ambiguous for a firm conclusion to be drawn either way. A low correlation between implicit and explicit measures can be interpreted in any number of ways; it could be evidence for a dissociation of attitudes; it could indicate an independence of measures; or it could even indicate a lack of convergent validity (Perugini, 2005). As has been pointed out

question that I advise being carried on in further research, for my real attitude may not necessarily always be the best predictor of my behaviour.

by psychologists and philosophers, what really matters to us when we investigate implicit attitudes is how predictive they may be of behaviour (Fazio & Olson, 2003; Perugini, 2005).

In a study conducted by Perugini (2005), strong evidence was found to support the dual attitudes model over the independence/additive model of attitudes. However, results also supported a third hypothesis that implicit attitudes and explicit attitudes may *interact together* to manifest behaviour, particularly when it is the case that implicit and explicit attitudes *align* (Perugini, 2005).[61] Overall the results from this study support the notion that implicit attitudes and explicit attitudes can *both* play a role in the production of behaviour. Combined with motivational factors, and the opportunity to deliberate and then override behaviours, either attitude could potentially determine behaviour.

Before continuing, I wish to express a very important point. Given that evidence can be found to support either model of implicit attitudes, and thus it remains (at least presently) a relatively open possibility for either model to be correct, for the sake of this project, I will be operating on the assumption that implicit and explicit attitudes are distinct constructs. However, this does not mean that the following project is only of importance or significance to those who endorse a dual attitudes model. Suppose that implicit and explicit attitudes are a single construct; we have one attitude, however the dissociation observed when we measure this attitude on implicit and then explicit scales implies that there may be dimensions of our single attitude that we are not aware of (or, at least, were not previously known to us). In this case, we might ask whether we are responsible for always acting on our attitudes, or perhaps, whether the existence of these other dimensions might make it such that we are not always responsible for the acts which issue from them. As such, whilst I will still proceed with this project on the premise that

[61] As Perugini (2000) rightly points out, a bias exists in the literature surrounding implicit attitudes which only asks how they operate *when there is conflict between implicit and explicit attitudes*. Perugini's study considers the effects that these attitudes have on behaviour when they are consistent with each other.

implicit attitudes are dual constructs (given the diversity of the evidence in favour of such a position), I wish to advise that this investigation is still of importance to those who do not necessarily endorse this model of attitudes.[62]

There are reasons for doing so; if it were the case that there was only a single attitudinal construct measured in two ways, then implicit attitudes don't appear to raise concerns for moral responsibility over and above the challenges already presented by attitudes more generally (or by explicit attitudes); the problem collapses. As this is the more philosophically interesting problem (considering implicit and explicit attitudes to be dual-constructs), and as it is an open possibility that this model holds true, I will be proceeding with this project on this assumption.

The conclusions I wish to draw from the work presented in this chapter is that implicit attitudes are an attitudinal construct which operate and exist (at least somewhat) outside of conscious awareness. They nonetheless can have a significant impact on our behaviour. While evidence is inconclusive as to whether implicit attitudes and explicit attitudes are a single or divergent construct, either way, I argue implicit attitudes have been demonstrated to have at least an interactive effect on behaviour with explicit attitudes. As implicit attitudes have been demonstrated to sometimes play a contributing role in the deliberative mechanism which issues potentially discriminatory or prejudicial behaviour, I believe that the behaviours they cause can, and should, be properly examined according to the Fischer and Ravizza (1998) framework outlined in chapter one.

[62] Whilst acceptance of my conclusions drawn from this project will not depend on which particular view of attitudes the reader adopts (be it the dual construct model or the single construct model), it *will* depend on exactly how that view of attitudes is constructed.

Moving forward, if implicit attitudes have this role in the deliberative mechanism, I ask to what extent is the deliberative mechanism which manifests our implicit attitudes reasons-responsive? As Fischer and Ravizza (1998) have outlined, we may only be held responsible for those behaviours of ours which issue from a reasons responsive deliberative mechanism. Given the nature of implicit attitudes, does the influence of implicit attitudes affect the reasons-responsiveness of the deliberative mechanism? I believe the answers to these questions lie in the degree to which implicit attitudes are reasons-responsive themselves. To solve this question, in the following chapter, I will determine to what extent implicit attitudes have an associative or a propositional structure.

Chapter Three – Implicit Attitudes: Beliefs or Associations?

In chapter one of this project I presented the theory proposed by Fischer and Ravizza (1998) that argued for a moderately reasons-responsive model of responsibility. I examined their claim that we can only be held morally responsible for those actions which issue from a deliberative mechanism which is capable of recognizing a coherent and consistent pattern of reasons, including some moral reasons, and which could respond to at least some of them if they were sufficient reasons to do so. If our actions are issued from a mechanism such as this, then Fischer and Ravizza argue that we can be held morally responsible for that action.

In chapter two I introduced the cognitive construct, implicit attitudes, and provided an inquiry in to their nature, structure, operation, and the testing procedures used to examine them. This inquiry revealed that implicit attitudes are known to play some sort of role in the production of behaviour. Specifically, implicit attitudes are *a part of* the deliberative mechanism which issues in behaviour; such as behaviour which manifests our implicit attitudes. However, it was established that the contribution implicit attitudes make to the mechanism is inconsistent. At times the implicit attitudes seem to function as reasons for acting, and respond to other semantic content. At other times implicit attitudes seem to be resistant to revision, other semantic content, and instruction.

In this chapter I will be attempting to determine to what extent implicit attitudes are reasons-responsive. As I go on to explain in chapter four, given that implicit attitudes are part of the deliberative mechanism which produces behaviour, it is important to establish the extent to which implicit attitudes are reasons-responsive. I propose that one way to determine whether implicit attitudes have these characteristics or not is to determine the extent to which they have an associative or propositional structure. In other words, should we describe implicit attitudes as ‘associations’ between concepts, or should we regard them as being akin to something like ‘beliefs’?

Determining which of these structures (if any) implicit attitudes have seems at least intuitively reasonable for a number of reasons. Firstly, an associative structure would more closely account for the automaticity, unresponsive, and sometimes biasing effect that implicit attitudes display than a propositional model. On the other hand, a propositional structure implies that implicit attitudes ought to be readily updatable and malleable when presented with new evidence or instructions. A propositional structure like this might suggest that implicit attitudes are reasons-responsive, which is a claim that seems at least harder to defend with an associative framework. And again, as I will go on to discuss in chapter four and my conclusion, this discussion also has implications for how we should go about managing or solving our issues of manifested implicit attitudes.

Drawing predominantly on the works of Mandelbaum (2015) and Levy (2014), I will present the evidence for and against both an associative and a propositional model of implicit attitudes. Firstly, I will present the reasons why implicit attitudes might be considered to have an associative structure. I will demonstrate that while the associative view is able to account for certain traits of implicit attitudes, it is not sufficient to explain their inconsistent reasons-responsiveness.

Secondly, I will provide evidence to support the notion that implicit attitudes have a propositional structure; demonstrating that factors such as inferential promiscuity, and implicit attitudes featuring as reasons for behaviour, can be explained by such a model. However, I will also show that implicit attitudes do not behave as consistently as bona fide beliefs, and that the propositional model is not sufficient to understand implicit attitudes.

Finally, having demonstrated that implicit attitudes are neither mere associations, nor entirely beliefs, I will be adopting and endorsing a third, 'middle-ground', alternative developed by Neil Levy (2014). I will be advocating that implicit attitudes are more appropriately described as 'patchy endorsements'. I argue that this alternative accounts for the inconsistent behaviour of implicit attitudes, and will show in chapter four that a patchy endorsements structure is compatible with Fischer and Ravizza's (1998) theory of responsibility.

Firstly, one reason for considering an associative model of implicit attitudes in the first place is because an extensive amount of the literature and research surrounding implicit attitudes seems to *assume* an associative structure. For instance, the foremost testing procedure for implicit attitudes, the Implicit *Association* Test, measures *associations* between concepts. Other sources cite implicit attitudes as reflecting "an *associative* system" marked by "slow processes of repeat pairings between an attitude object and related evaluations… and affected by *associative* information" (Rydell & McConnell, 2006, p. 995). Dasgupta and Greenwald (2001) remark about how response facilitation is "interpreted as a measure of the *strength of association* between object and evaluation". As an associative structure appears to be such a widely held assumption, such a structure merits considerable investigation. If we are going to make assumptions about a theory, then we ought to validate those assumptions. Such is one of my goals now.

Associations are automatically activated links between representations in the mind (Hughes, Barnes-Holmes & De Houwer, 2011). They are typically formed automatically and passively through repeat exposure to consistent pairings of stimuli in the environment; a conditioning paradigm, or a product of learning (De Houwer, 2011; Levy, 2014). Typically, associations are formed in at least one of three ways: the concepts are encountered at the same time (contiguity)[63]; the concepts may resemble each other (resemblance)[64]; or the associations can be formed on the basis of a cause and effect relation (causality)[65] (Mandelbaum, 2015). In each of these cases, an agent may come to develop an associative connection between attitude objects and attitudes. For example, I may come to associate 'black people' with 'danger', because black people are often paired with stimuli implicating them as dangerous[66].

One aspect of associative structures is that they are typically automatically activated in response to the perceptual experience of a stimuli (Hughes, Barnes-Holmes & De Houwer, 2011). Thus, my concept of 'danger' may be triggered upon merely seeing a black person. Moreover, associations are bi-directional, such that activating one concept ought to activate the other, and vice versa (Mandelbaum, 2015). For instance, activating the concept, 'black person', ought to activate my concept of 'danger'.[67] Reversing this, if my concept of 'danger' is activated, one response ought to be the activation of my concept of 'black people'[68]. Two features of associations come out of this bi-directional component; (1) they are believed to adhere to a 'spreading activation' model, and (2) that associations contain no relational information between concepts. I move to discuss each of these features.

[63] An example of an association formed through contiguity would be that I associate thunder with lightning because they are frequently found together.

[64] An example of an association formed through resemblance would be that I associate the picture of Sydney (city) on my postcard with the city of Sydney, because these two stimuli share a visual resemblance. Resemblance is also not limited to visual resemblance. For example, stimuli can also have an aural resemblance, such as associating trucks with buses because they sound similar.

[65] An example of an association formed through cause and effect would be I associate smoke with fire because fires cause smoke.

[66] Such as crime rates, news reports, frequent pairings of 'black people' and 'danger' in conversation, etc.

[67] Another example of this could be 'salt' and 'pepper'

[68] There are of course many different concepts that may be activated in response to the activation of a single concept. Activating the concept, 'danger', may also activate my concept of 'violence', 'death', 'snakes', etc. However at least one of these responses would be activation of the concept 'black people' if one associates black people with danger.

Spreading activation models postulate that concepts are retained in memory as 'nodes'[69], and that a collection of nodes form an 'associative network'. Activation 'spreading' is the idea that activating one node will activate other nodes in the network, and, based on their proximity to the initial activation, certain nodes may be activated faster (or more readily) than others (Mandelbaum, 2015).[70] That is, the more frequently two concepts are activated, the quicker they may be activated together, and the more readily we may associate them. Furthermore, depending on the strength of the connections between nodes, and because each node can have many connections, activating one node can *prime* (or 'dispose') the agent to respond in certain ways instead of others[71]. As a result, if implicit attitudes have an associative structure then they may serve to dispose agents towards particular kinds of behaviour; they may serve a biasing effect.

Associations are symmetrical (bi-directional) because they do not contain any internal representational content which describes the nature of the relation between concepts (Hughes, Barnes-Holmes & De Houwer, 2011; Mandelbaum, 2015). In other words, the concepts 'black person' and 'danger' are activated together regardless of what kind of relation that is; 'black people *are* dangerous', or, 'black people *are not* dangerous'. This lack of a satisfaction condition means that associations are not formed on the basis of logic and reason; they are merely products of conditioning (Levy, 2014). Being products of conditioning, we ought to be able to change or

[69] A 'node' may be thought of as a packet of mental content. For example, I may have a 'school' node, or a 'work' node.

[70] Let's take an example. As above, I may have a 'school' node, but I may also have a collection of other nodes in the network, like a 'playground' node, one for 'classroom', and one for 'pencil'. Spreading activation postulates that activating my school node would also activate these other nodes in the network. Furthermore, the stronger the connections are between each node, the quicker I will be to report the association between those two nodes. And this is based on how frequently those two nodes are paired together. Therefore, I may more readily associate 'school' with 'classroom' because those two concepts are more frequently paired together than 'school' and 'pencil'' (which can be paired more frequently with other concepts, like 'pen').

[71] For instance, because I associate 'black people' with 'danger', if this connection is particularly strong, then upon presentation of the 'black person' stimulus, I may be disposed to respond with fear instead of happiness or excitement.

remove associations through extinction procedures[72] or via counterconditioning[73]. One way to remove my association between ‘dangerous’ and ‘black people’, for example, may be to start exposing myself to instances of particularly helpful black people as opposed to dangerous ones (counterconditioning). Thus associations ought to be moderated by manipulating the factors affecting the strength of the association (De Houwer, 2014).

Considering these traits of an associative view then we can understand why we may regard implicit attitudes as adhering to this structure. As I showed in chapter two, there are instances where implicit attitudes seem to be formed automatically and passively, without any conscious reasoning behind them. An associative model also accounts for the aforementioned automatic activation that implicit attitudes display, as outlined in chapter two. Furthermore, we can observe bi-directional relations in implicit attitudes, and there is research which supports the spreading activation models of mental cognition.

However, in a number of ways, an associative account falls short of fully explaining the nature of implicit attitudes. Firstly, if implicit attitudes were formed entirely on the basis of conditioning exercises then everyone exposed to the same environment ought to possess the same implicit attitudes, and this is not the case (Holroyd, 2012). An agent’s implicit attitudes seem to be determined by more than just their environment. Implicit attitudes seem to be shaped by also their beliefs and values and motivations (Holroyd, 2012; Holroyd & Kelly, forthcoming; Hofmann, Gschwendner, Castelli & Schmitt, 2008) among other things.

Furthermore, while it is true that implicit attitudes do show a certain degree of enduring stability across time and place (Fazio & Olson, 2003; Levy, 2014), as associations should, other research has also demonstrated that implicit attitudes *can* be subject to change and revision if the

[72] Extinction is the process of presenting one relata without the other until the association is broken

[73] Counterconditioning is the process of changing the valence of the relata until the opposite (or different) association is formed.

right motivation is present, and the right techniques are applied (Holroyd, 2012; Holroyd & Kelly, forthcoming). Examples of these techniques include extinction or counterconditioning procedures. However, even these procedures which should moderate our implicit attitudes, do not always do so consistently. For instance, some studies have shown that applying these procedures can often produce short-term results, and the subject will return to the initial association after a short time (Levy, 2014; Mandelbaum, 2015).

What these pieces of evidence suggest is that implicit attitudes do not perfectly adhere to an associative model. While they demonstrate certain associative traits, they are not merely associations. The associative model therefore does not appear to accurately describe the behaviour of implicit attitudes. However, alternative models have been suggested. I turn now to consider the propositional model, and determine whether it is able to account for the nature of implicit attitudes.

The propositional model posits that implicit attitudes may be akin to propositional structures like beliefs. One of the characteristics of beliefs which makes them different from associations is that they have what are referred to as 'satisfaction conditions'[74]. Therefore, attitudes adhering to a propositional structure must have their constituents stand in a specific relation to one another (De Houwer, 2014). Unlike associations, which merely connect two relata, beliefs (and propositionally structured devices) specify the nature of the relation between relata.[75] In other words, there is a particular way that these concepts are related. Moreover, beliefs are truth apt[76] (Mandelbaum, 2015) and are pragmatic and context-relative (Frankish, 2015).

[74] A satisfaction condition states that the conditions of the proposition are only satisfied if its intentional objects stand in a particular relation to one another (Levy, 2014).

[75] For example, propositions specify that 'black people *are* dangerous', as opposed to merely associating 'black people' and 'dangerous'.

[76] Meaning that beliefs have a truth value

While associative structures are a product of conditioning (whereby relations are formed based on frequent pairings of stimuli together), propositional structures like beliefs can be formed through instruction, reasoning, and single instances of inference (Hughes, Barnes-Holmes & De Houwer, 2011). As is discussed by Levy (2014), and mentioned by Mandelbaum (2015), there are two primary characteristics of beliefs. Firstly, beliefs are '*inferentially promiscuous*'. This means that our beliefs are able to interact with other beliefs and mental states, and so affect behaviour. For example, if my implicit attitude were propositionally structured, my belief that black people are dangerous may interact with my belief that I would be safer walking on the other side of the street, and then lead me to do so.

Secondly, beliefs update according to new evidence, and through the interaction with other mental states and semantic content. Thus, beliefs are also *responsive to evidence* (Levy, 2014). What makes beliefs different from associations is that beliefs are malleable in the face of reason and evidence (De Houwer, 2014; Hughes, Barnes-Holmes & De Houwer, 2011; Levy, 2014; Mandelbaum, 2015). Propositional structures can be changed and updated in response to reason, meaning that beliefs can be argued against and revised if sufficient reasons are presented. Incidentally, this means that beliefs can feature as premises in reasoning; perhaps as reasons for behaviour, or for being disposed to acting certain ways (Levy, 2014; Mandelbaum, 2015). If my implicit attitude about black people and danger is a belief (or at least propositionally structured) then I ought to be able to change this attitude by arguing or rationalizing it away. Perhaps I read about significant examples to the contrary (Blair, 2002), or read studies indicating that there are no racial differences to be afraid of.

Finally, Frankish (2015) argues that "explicitly believing that *p* is to take *p* as a premise in one's reasoning and decision making". He goes on to say that, by accepting a proposition, we commit ourselves to stand by its truth; we hold it to be true, and will act in accordance with that proposition on the assumption that it is true. Levy (2014) contributes a similar opinion, that if my implicit attitude is propositionally structured then I endorse that particular relation between its constituents. That is, I hold them to stand in a determinate relation. If my implicit attitude is so structured, then I commit myself to *believing* that black people are dangerous; I take this relation to be true and will act as if it were true.

Together, these factors construct an image of implicit attitudes that is admittedly more flexible and malleable than the associative view. On the propositional account, implicit attitudes have specific relations and are not merely products of learning. As a result, implicit attitudes ought to be able to interact with other mental states and feature as reasons in the production of behaviour, and as a premise in reasoning.[77] These are traits which I have demonstrated to be ascribable to implicit attitudes. Moreover, there are studies which demonstrate that implicit attitudes are responsive to rationalization and argumentation in that they can adapt to the presentation of evidence (Mandelbaum, 2015).

The propositional model can even account for the automaticity of implicit attitudes[78], as once propositional structures have been formed, they can be automatically retrieved from memory because “they’re encoded as episodic-like representations” (De Houwer, 2014). In other words, once propositional structures have been formed, they are stored in memory and be still be automatically retrieved. This would allow the propositional model to account for the automaticity displayed by implicit attitudes. Thus the propositional model does a lot of work in explaining

[77] I wish to clarify that it is not the case that implicit attitudes can only interact with other mental states if they are propositionally structured. Associatively structured implicit attitudes are also able to interact with other mental states and behaviours, but they each do so in different ways. Associations, as I have said, may exert a biasing influence on behaviour or deliberation. Propositionally structured attitudes respond to the other mental states and semantic content in a reasons-responsive way, and are capable of featuring in inferences. Thus, when I talk here about propositionally structured implicit attitudes being able to interact with other mental states, what I am referring to is their reasons-responsive interaction, which associations cannot do.

[78] A characteristic shared by the associative model.

implicit attitudes. I would argue it even manages to do some of the work provided by an associative model.

However, despite the evidence above, there are still reasons to keep us from endorsing the propositional model as the most viable account of implicit attitudes. Firstly, while implicit attitudes have been found to respond to reason at times, the effects of rationalization can temporary, after which the attitude often reverts back to its original state (Levy, 2014). Secondly, and more to the point, implicit attitudes sometimes do not update consistently the way beliefs do (if at all). Payne, Cheng, Govorun & Stewart (2005) demonstrated that subjects' implicit attitudes were often immune to instructions; that their behaviour did not take account of the information provided by the experimenters. If implicit attitudes were indeed propositionally structured in the same way as bona fide beliefs, then these attitudes ought to respond to the instructions. These are just a couple of reasons as to why the propositional model falls short as an account of implicit attitudes.

When we consider these different structures from the perspective of the Fischer and Ravizza (1998) moral theory, it seems the most desirable outcome would be that implicit attitudes be propositionally structured. A propositional structure does a lot of work for this theory. It can account for the automaticity observed in implicit attitudes; it can explain how it is that implicit attitudes may exist and function outside of consciousness[79]; and it also has predictive and heuristic worth (De Houwer, 2014).

As I go on to discuss at length in chapter four, the theory of moral responsibility that Fischer and Ravizza (1998) are positing seems to require that implicit attitudes possess a degree

[79] Mandelbaum (2015) argues that the cognitive structure which underpins implicit attitudes, even on a propositional account, is unconscious. We needn't believe that the *valences* attached to beliefs must be unconscious; in fact, as Mandelbaum reasons, the valences may be conscious even if it is the case that the relata in question are not conscious.

of reasons-responsiveness. A propositional model provides us with this characteristic. While it may be the case that implicit attitudes do not update as regularly or as consistently as bona fide beliefs do, a little insensitivity to reasons is not especially condemning.[80] We are human after all, and are not going to recognize every piece of evidence presented to us, nor always update our beliefs in a rational way. Certainly there seem to be plenty of beliefs which seem irrational, or at least, do not appear to update rationally in the face of evidence to the contrary.

What is important is that, if implicit attitudes are beliefs, then there is little reason to assume they might be any less reasons-responsive than conscious beliefs (Levy, 2014). We can, and do, hold people morally responsible for actions they perform which are the product of incorrect conscious beliefs; particularly if the moral agent has the ability to recognize and consider evidence, and revise said belief. This simply does not seem appropriate on an associative model, as the implicit attitude would be resistant to change, merely a product of learning, and could merely dispose a person towards action, rather than feature as a premise for acting. It therefore does not seem appropriate (on the Fischer and Ravizza (1998) framework) to hold agents morally responsible for actions which issue from a deliberative mechanism which is not responsive to evidence in the way dictated by the associative model.

As I have already discussed, implicit attitudes seem to be more than mere associations, yet are not quite as consistent in their behaviour as bona fide beliefs. It would appear implicit attitudes adhere to neither an associative nor a propositional structure. At this point I argue for a third, ‘middle-ground’ alternative put inspired by Levy (2014), that implicit attitudes “are a sui generis state” called ‘patchy endorsements’. ‘Endorsements’ are a mental construct which have propositional structure, entail satisfaction conditions, and commit an agent to taking the world to

[80] In fact, Fischer and Ravizza’s (1998) theory actually makes a similar specification. It is not necessary that the deliberative mechanism be able to recognize and respond to *every* sufficient reason presented to it; such would be too stringent a requirement.

be a particular way; these being characteristic shared by implicit attitudes. However, they are described as 'patchy' because they only feature in *some* kinds of inferences, and respond to *some* pieces of evidence that we would expect bona fide beliefs to respond to (Levy, 2014). In other words, implicit attitudes are not as consistent as beliefs, but are certainly more reasons-responsive than mere associations would allow.

Thus by considering implicit attitudes to be patchy endorsements, rather than strictly associations or beliefs, we manage to provide an account which caters for the inconsistencies in the behaviour of implicit attitudes. Implicit attitudes have propositional structure insofar as they *can* be reasons-responsive[81]; they have satisfaction conditions, which means that they commit the individual to taking the world to be a certain way; and we manage to retain the aspects of the associative model which explains why they remain consistent throughout time and different contexts. I argue that considering implicit attitudes to be patchy endorsements does a lot of work in properly explaining implicit attitudes.

The most significant benefit of a patchy endorsements account, however, is that it entails that implicit attitudes *can* be reasons-responsive to *some* extent. This is an important possibility for the Fischer and Ravizza (1998) account, as I will demonstrate shortly in chapter four. This feature of implicit attitudes ties together with my comments in chapter one that responsiveness and moral responsibility are matters of degrees. Fortunately, as a result of their 'patchy' nature, both moral responsibility, and the reasons-responsiveness of implicit attitudes can be discussed in terms of degrees. However, given this inconsistency, Levy (2014) argues that we must examine each behaviour on a case by case basis to determine when (and whether) our implicit attitudes are interacting with semantic content in a reasons-responsive way, and *to what extent they may be*

[81] Implicit attitudes are not required to always be reasons-responsive, however the possibility exists.

doing so. Even so, it is clear that even patchy endorsements meet the requirements and standards for talking about moral responsibility that I discussed in chapter one.

I will proceed now to demonstrate that such an approach to implicit attitudes is of particular importance to the Fischer and Ravizza (1998) account of responsibility. In chapter four, I will argue that if we wish to determine the extent to which someone is morally responsible for their actions which manifest (and have been influenced by) our implicit attitudes, then we need to be able to determine the extent (and nature) of their contribution to the deliberative mechanism.

Chapter Four – Fischer and Ravizza on Implicit Attitudes

In this chapter I will be examining how Fischer and Ravizza (1998) should regard actions which manifest our implicit attitudes; ought we to be held morally responsible for such actions? To answer this, I will be addressing each criterion for responsibility set out by Fischer and Ravizza in chapter one, and determining whether behaviours manifesting implicit attitudes satisfy these requirements, given what we understand of implicit attitudes from chapters two and three.

First, I will address the notion of reasons-responsiveness. In chapter two I demonstrated that implicit attitudes do not appear to respond to reasons consistently. How does this trait of implicit attitudes affect the operation of the deliberative mechanism which produces behaviour, and to what extent is the deliberative mechanism reasons-responsive?

Secondly, I will outline Fischer and Ravizza's (1998) 'tracing' argument, and determine whether this may be suitable grounds for holding an agent as morally responsible. I will argue that, even if our implicit attitudes are not reasons-responsive at the right time, that we might still be held morally responsible in lieu of our ability to exercise what Holroyd & Kelly (2012) refer to as 'ecological control'. I will suggest that we can use this argument to improve the consistency in the reasons-responsiveness of our implicit attitudes.

Finally, I will conclude this chapter by determining whether implicit attitudes meet Fischer and Ravizza's (1998) requirements for guidance control. To do this, I will be discussing the concept of 'irresistible psychological impulses'. In chapter one I briefly mentioned that agents acting under the influence of genuinely irresistible psychological impulses may be absolved of moral responsibility because it undermines their guidance control. In this chapter I will be extending upon this thought, and determining whether we ought to consider implicit attitudes to be irresistible psychological impulses, and whether this affects how we would ascribe moral responsibility.

In chapter one of this project I set out the requirements for the Fischer and Ravizza (1998) theory of moral responsibility. Having posited the notion of reasons-responsiveness, Fischer and Ravizza contend that an agent is morally responsible for the actions he performs which issue from his own reasons-responsive deliberative mechanism. In other words, the mechanism which produces a given action must be moderately receptive to a range of reasons for acting, including at least some moral reasons, and must also be capable of responding to at least some of these reasons (as well as being the agent's own). Therefore, if, when an agent acts, the deliberative mechanism was capable of recognizing a consistent pattern of reasons for acting, and would have responded to at least some reasons to do otherwise, the agent is morally responsible for their action.

In chapter two I introduced implicit attitudes. I determined that implicit attitudes are a type of mental evaluation which we hold in regards to objects and ideas in the world. I determined that while in some instances these attitudes may be unconscious or operate outside of conscious awareness, what sets them apart from our explicit attitudes is the fact that implicit attitudes are typically automatically activated. Implicit attitudes can have an influence on our behaviour. In particular, the influence of negative implicit attitudes can result in morally negative behaviours.[82] I also began to highlight that implicit attitudes seem at times to be resistant to change and evidence, and yet at other times respond to other reasons, evidence, and mental states. Because the influence of implicit attitudes can lead to morally relevant behaviours, and because of the perceived inconsistency in their operation, I sought to apply the Fischer and Ravizza (1998) theory to the behaviours which manifest our implicit attitudes, and determine to what extent we are morally responsible for such behaviours.

[82] Such negative behaviours include, but are not limited to, rudeness, discrimination, and prejudice

The focus of this study has been to what extent are we morally responsible for those behaviours which manifest our implicit attitudes. In order for us to ascribe moral responsibility, the deliberative mechanism issuing that behaviour must be capable of recognising a pattern of reasons and how they relate to each other. However, it does not always exhibit this consistent and coherent pattern of reasons-responsiveness; there are times when the mechanism is unable to recognise reasons, or even respond to them. This can be the result of a number of contributing factors. As I have argued in the preceding chapters, implicit attitudes are one such contributing factor. Our implicit attitudes comprise a component of the deliberative mechanism itself, and because of this they can affect and shape behaviour, and even become manifest in behaviour.

Because implicit attitudes are a part of the deliberative mechanism, I argue that the deliberative mechanism as a whole is (to some extent) made more or less reasons-responsive in virtue of the degree of reasons-responsiveness of our implicit attitudes. That is, the degree to which our implicit attitudes are reasons-responsive contributes to the degree of reasons-responsiveness of the deliberative mechanism overall; if our implicit attitudes are particularly reasons-responsive, then the mechanism is made more reasons-responsive, and so we may ascribe (possibly a greater degree of) moral responsibility for actions which issue from this mechanism. If, however, our implicit attitudes are not particularly reasons-responsive, then the mechanism is made less reasons-responsive, and it becomes harder to ascribe significant moral responsibility to the agent[83]. Therefore, when we are considering behaviours which are influenced by our implicit attitudes, and which even manifest those attitudes, we need to determine the extent to which the

[83] This does not mean it is impossible to ascribe moral responsibility if the implicit attitude is not reasons-responsive. It could be the case that the agent is simply *less* morally responsible for the action than he otherwise would have been, had his implicit attitude been more reasons-responsive. Remembering that in chapters one and three I raised the notion that moral responsibility should be thought of in terms of degrees, so too should we be thinking of the reasons-responsiveness of the implicit attitudes as a matter of degrees. In other words, patchy implicit attitudes which are significantly reasons-responsive may contribute towards making the agent *more* responsible for the actions issuing from their deliberative mechanism than implicit attitudes which are less reasons-responsive. The agent may still be regarded as being morally responsible in both scenarios, however not to the same extent.

implicit attitudes, at that time, were reasons-responsive. This served as the motivations for chapter three.

In chapter three I examined the reasons-responsiveness of implicit attitudes by determining whether they have an associative or propositional structure. The purpose of this was to determine the extent to which implicit attitudes are responsive to reason, and account for the inconsistency in their receptivity outlined in chapter two. For instance, if implicit attitudes had an associative structure, then they would not be very receptive to reasons, they would not update regularly in response to evidence. Because of this, their influence on behaviour would serve more of a biasing effect than a deliberative effect.[84]

If implicit attitudes had a propositional structure, however, they would be able to interact with other mental states and processes in a reasons-responsive way, thereby updating and changing in response to evidence. Propositionally structured attitudes would not have a biasing effect on behaviour, but would rather respond to reasons for acting one way instead of another based on the way they interact with other pieces of evidence.

It was found that implicit attitudes fit neither of these structures and instead were more appropriately described as patchy endorsements. In this sense, implicit attitudes are responsive to reason but not as consistently, or to the same extent, as bona fide beliefs, but are also much more responsive to reason than mere associations. However, by describing them as ‘patchy’, we acknowledge that the degree to which an implicit attitude is reasons-responsive will vary across context and situation. This in turn may account for why it is the deliberative mechanism may not be consistently reasons-responsive in regards to behaviours concerning implicit attitudes.

[84] Remember that if the implicit attitude has an associative structure, then in virtue of its associative structure, the implicit attitude interacts with other mental content and reasons by disposing the agent to act one way instead of another. Being associative in structure does not stop the implicit attitude from interacting with other semantic content, it just keeps it from interacting in a reasons-responsive way.

Returning to the discussion above, if it is the case that the implicit attitude is capable of responding to other reasons, then, in virtue of this, the resulting behaviour issues from a mechanism that was made more capable of considering a variety of reasons. If the mechanism was indeed capable of recognizing other reasons for acting, could have responded to another sufficient reason, but failed to do so, then we may hold the agent morally responsible.[85] Thus, the mechanism is at least made more reasons-responsive in virtue of the implicit attitude being reasons-responsive, and the agent may be held morally responsible for the behaviour which manifests that implicit attitude.

Alternatively, the implicit attitude influencing the mechanism may not be properly reasons-responsive. That is, it might have hindered the ability of the mechanism to interact in a reasons-responsive way with other reasons and factors. If the implicit attitude is not reasons-responsive, then it may be functioning more ‘associatively’ and contributing a biasing effect in the mechanism. If this is the case, while the mechanism may able to recognize a variety of reasons for acting, its deliberative processes cannot be properly reasons-responsive because the implicit attitude is not inferentially promiscuous; it is not interacting with the other reasons in the right way, and is instead disposing the mechanism towards one course of action. If the mechanism functions this way, thereby manifesting the implicit attitude in behaviour, then it does not seem appropriate to hold the agent as morally responsible as we would have, had the attitude been reasons-responsive, because the mechanism was not properly reasons-responsive. In this instance, the mechanism was not able to *really* recognize other reasons for acting, and because of the biasing influence of the implicit attitude, it was not given the opportunity to respond. Again,

[85] There could be a few reasons that the mechanism did not to react to the other presented reasons. Perhaps the mechanism devalued these other reasons; perhaps they were not deemed as sufficient as other reasons and were then decided against. I am sure there are other possibilities as to why other reasons may have been overlooked.

the mechanism appears to be more or less responsive to reason in virtue of the reasons-responsiveness of the implicit attitude.

What I hope to have shown is that the degree to which the mechanism is reasons-responsive is in virtue of the degree to which the contributing implicit attitude is reasons-responsive. If the implicit attitude is reasons-responsive, and can properly interact with other reasons, then the mechanism is made more or less capable of recognizing these other reasons and considering all possibilities. If this is the case, then the reasons-responsiveness criteria set out by Fischer and Ravizza (1998) is satisfied and we may hold the agent morally responsible. If the implicit attitude is not reasons-responsive, and therefore contributes a biasing effect, then the mechanism cannot be said to have been able to properly recognize other reasons for acting outside of the implicit attitude. So it does not seem appropriate to say the agent is morally responsible for manifesting the implicit attitude.

Therefore, to determine the degree to which someone is morally responsible for manifesting their implicit attitude, we need to determine how reasons-responsive the implicit attitude was at that moment. At the end of chapter three I noted a suggestion by Levy (2014) that, in order to determine when and whether an implicit attitude is reasons-responsive, and interacting with other semantic content, we must consider the influence of implicit attitudes on a case by case basis. In light of the arguments I have presented above, I support this move.

Fischer and Ravizza (1998) specify that a mechanism must be capable of recognizing a coherent and consistent pattern of reasons, as well as being capable of responding to some reasons. When our actions manifest our implicit attitudes, in order to determine the degree to which the mechanism is capable of this, we must examine each action on a case by case basis.

However, it nonetheless remains the case that implicit attitudes *can* be reasons-responsive[86], and, in cases which we manifest our implicit attitudes, we may be held morally responsible. Certain behaviours then *can* satisfy this criterion for moral responsibility.

I turn now to consider a different point. Suppose that the implicit attitudes acting at the time are not reasons-responsive. While Fischer and Ravizza (1998) contend that if the mechanism is not reasons-responsive then we cannot hold the agent morally responsible, they posit the idea to examine the causal history of that action[87]. Depending on the causal history of the action then we may still be inclined to say that an agent is morally responsible even if his implicit attitudes at the time were not reasons-responsive.

Suppose an agent performs an action at time T2, and does so in a manner that issues from his implicit attitudes when they are not reasons-responsive. We need to ask whether the agent is responsible for his implicit attitude *not* being reasons-responsive *at that time*. That is, has he, before this, at time T1, acted in such a way as to make his implicit attitudes non-responsive at time T2? Fischer and Ravizza posit a 'tracing' argument which states, "when one acts from a reasons-responsive mechanism at T1, and one can reasonably be expected to know that so acting will (or may) lead to acting from an unresponsive mechanism later at T2, one can be held morally responsible for the actions performed at T2" (Fischer & Ravizza, 1998, p. 50).[88] In other words, we may say that the agent is to some extent morally responsible for his actions manifesting non-

[86] This does not mean that they always are, it just means that there is the possibility that in some scenarios the implicit attitudes may be reasons-responsive.

[87] By 'causal history' Fischer and Ravizza (1998) refer to the events, circumstances, actions, and decisions leading up to, and contributing, to that action.

[88] As an example, suppose I go to the local pub and have too much to drink and become intoxicated. This is T1. I then decide to drive home, and, because of my intoxication, I swerve in my lane and cause a dead-on collision with another vehicle. This is T2. At time T2 I am intoxicated and so am not acting from a reasons-responsive deliberative mechanism. However, Fischer and Ravizza's (1998) tracing argument says that I am morally responsible for the car crash because I am responsible for putting myself in an intoxicated state. While I am not reasons-responsive when I crashed the car, I was reasons-responsive earlier when I decided to drink myself in to that state. Because of this, I am morally responsible for crashing the car.

responsive implicit attitudes *if* he could have acted in such a way *before* the action that would have prevented his implicit attitudes from being non-responsive.

What I argue is this: even though the mechanism may be made less reasons-responsive in virtue of his implicit attitudes being non-responsive, the agent still retains a degree of control over the causal history of those attitudes which makes him morally responsible in lieu of the tracing argument. I believe Jules Holroyd (2012) makes an argument for control along similar lines. Holroyd argues that while we may lack the *direct and immediate* control over our implicit attitudes, we do have *indirect, long range* control "for the end goal". What this means is that, while we may not have control over the attitude itself, we can still exercise control over the intermediate steps and processes which affect the reasons-responsiveness of our implicit attitudes, and these steps will indirectly affect the role implicit attitudes play in the mechanism. This line of argument is similar to Fischer and Ravizza's tracing argument over causal history.

Holroyd and Kelly (forthcoming) employ a model of 'ecological control' in regards to this argument. Ecological control is exercising actions without deliberative or reflective control (Holroyd & Kelly, forthcoming). We *take* ecological control when we reflectively decide to manipulate our mental states and environment in an effort to better shape and coordinate the cognitive processes which guide our behaviour (Holroyd & Kelly, forthcoming). The hope is that, if we manipulate our environment in a particular way (say by making certain features more salient), then we can indirectly manipulate our behaviour and make certain responses automatic, or change and shape our currently automatic responses.[89]

[89] For instance, I may have a fear of spiders, and I may automatically experience a sensation of fear when I am exposed to a spider or see one. A fear response is an automatic response, and I have no direct and immediate control over experiencing this response. However, by taking ecological control, by reflectively and consciously manipulating my environment, I may be able to change the way I automatically respond to spiders. For example, I may decide to surround myself with pictures of spiders so that I am more exposed to harmless spider stimuli; I may decide to read more about spiders to better understand them; or I may recite a mantra whenever I see a spider to discourage a fear response ("It's more afraid of me than I am of it", for instance). By doing this, I encourage certain

Holroyd & Kelly (forthcoming) believe such a methodology can be applied to implicit attitudes, and so we can take steps to mitigate the influence of negative implicit attitudes, or change the way they automatically respond to stimuli. Furthermore, that we can exercise ecological control in a bid to better enable our implicit attitudes to recognize and react to our other beliefs, values, and semantic content. By exercising ecological control over our implicit attitudes, we ought to be able to change the way they behave and increase their reasons-responsiveness, or make it more likely that they will be reasons-responsive. If we can do this, then we may hold agents morally responsible to some extent for manifesting their implicit attitude, even if those implicit attitudes are not reasons-responsive (as in the example above).

There are many ways to achieve this. To name a few, some implicit attitudes can be counteracted by exposure to contrary exemplars (Blair, 2002). Other evidence has found that agents who are committed to particularly egalitarian goals are also better able to suppress the expression of prejudicial implicit attitudes (Holroyd, & Kelly, forthcoming). In this case, an agent may exercise ecological control over his implicit attitudes by forming and pursuing certain values. Holroyd also suggests the use of 'implementation intentions'[90] (Webb, Sheeran & Pepper, 2010, as cited in Holroyd, 2012) as a way to shape our implicit attitudes and the way they interact with semantic content.

If our exercising of ecological control in such ways as described above are successful, then the newly formed responses and behaviours ought to bring our automatic judgments and emotional reactions (our implicit attitudes) in to line with our more explicitly held attitudes and

behaviours to become *routine*, and discourage certain behaviours. Eventually I ought to develop a new response to spiders, such as curiosity, or at least inhibited my fear response, based on the way I shape my environment.

[90] Implementation intentions are attempts to change behaviour "in the presence of a built-in conditional or environmental cue" (Holroyd, 2010). An example of this would be similar to the mantra from the earlier spider example (see Chapter Four, footnote 8); an agent who thinks black people are dangerous may say, "whenever I see a black person, I will think 'safe'". After enough recitations and implementation of the action, the implementation intention will become the new automatic response to the 'environmental cue' of seeing a black person.

intentions (Holroyd & Kelly, forthcoming).[91] In this way we exercise long range and indirect control over the manifestation of our implicit attitudes, and the way they interact with other reasons in the deliberative mechanism. Applied to the tracing argument, by shaping our environment and exercising ecological control at time T1 we can shape the way our reasons-responsive and non-responsive, implicit attitudes are expressed at time T2. Thus even if we encounter a scenario whereby our implicit attitudes are not currently reasons-responsive, if the agent is aware of the presence and potential influence of their implicit attitudes, we may still hold the agent morally responsible to the extent that he failed to exercise ecological control beforehand when they were reasons-responsive.[92]

Thus, in response to Fischer and Ravizza's (1998) requirement of 'reasons-responsiveness', the deliberative mechanism is made more or less reasons-responsive in virtue of how reasons-responsive are the attitudes which comprise it. When the deliberative mechanism manifests our implicit attitudes, I believe we need to consider on a case by case basis the degree to which the implicit attitude was reasons-responsive. Should we find that the implicit attitude is not reasons-responsive, then the agent may not be eligible for moral responsibility.[93] But if the agent is aware of the attitude and its potentially biasing influence, then the agent may be able to exercise ecological control prior to its manifestation in behaviour. [94] If the agent fails to exercise ecological control despite the ability and opportunity to do so, then we may say he is still morally responsible to some extent.

[91] In this regard, we have not necessarily changed the attitude, more how the attitude functions in the deliberative mechanism. We can develop the implicit attitudes in a reasons-responsive way; to be more reasons-responsive. This is the tracing element; we could have made the implicit attitudes more reasons-responsive.

[92] As I will discuss in my concluding chapter, this may place the requirement on the agent to discover and learn about their implicit attitudes, and this does bring with it its own implications for moral responsibility. I will discuss this later in my conclusion.

[93] We may more precisely say that, if the implicit attitude is not reasons-responsive, then the agent might simply be *less* morally responsible for the behaviour, instead of being absolved of responsibility entirely

[94] If the agent does exercise ecological control, and yet still manifests the behaviour, if it is found that the attitude was still not reasons-responsive, then we ought to not hold the agent morally responsible for the behaviour.

I have argued that we need to determine the reasons-responsiveness of our implicit attitudes on a case by case basis. I have also argued that this is both plausible and practical on the account developed in this project. To demonstrate how these conclusions might play out in practice, I return to discuss the case of the sexist job recruiter I mentioned earlier on in this project. Remember that the job recruiter is presented with a male and a (more qualified)[95] female candidate for a job, and yet he still chooses to hire the male. The recruiter still maintains however that he believes everyone should be treated equally, and he did not choose based on any sort of prejudice against women. I ask how would we evaluate the responsiveness of the deliberative mechanism in this scenario, and determine to what extent the recruiter is morally responsible.

First of all, the fact that the recruiter's actions are not consistent with his explicit beliefs or the facts about the females qualifications suggests that the recruiter must have acted on the basis of a deliberative mechanism that is not properly reasons-responsive.[96] Moreover, it is very likely that the ma must have been influenced by an implicit attitude that was not especially reasons-responsive, as this is the only explanation for his selection.[97] This discrepancy in his actions and his explicit attitudes ought to prompt us to determine whether he has an implicit attitude or not. This would require the agent to take an implicit attitudes test.[98] At this point we

[95] In this case, by 'more qualified' I also mean that the female is both more eligible for the job, better suited for the job, has superior qualifications for the job, and is generally the best option.

[96] If they were, he ought to have selected the female candidate. In all possible regards, she is the best candidate for the job; there is no conceivable reason why the male should be hired over her.

[97] I believe the only other explanation for this seemingly irrational action is that the recruiter acted completely randomly and irrationally. However, in this instance we are working with a typical functioning agent and so I will not be considering this possibility.

[98] While some may argue that having to take an implicit attitudes test makes things more complicated than practical, I argue the following: Firstly, the complexity of a practice ought not to affect the practicality have said practice. Sure, it may require more steps to implement, but even so the practice itself is still applicable in the real world. Secondly, a variety of credible implicit attitude tests are now readily available online and are easily accessible by the public. While these may not be as perfectly refined in their implementation as those which may be administered by psychologists, I argue that they are sufficiently accurate such that they are a valid option. Given their ease of use, their brevity, their accessibility, I believe that asking an employee to sit one of these tests is not too complicated a process.

will know whether the agent has an implicit prejudice against women or not. This has a few implications.

Firstly, at the moment when the recruiter failed to select the female candidate, I believe it is reasonable to say that his implicit attitude was not especially reasons-responsive; if it were then it would have been better able to interact with other incoming evidence, such as his explicit attitudes, and the superior skillset of the female candidate. In this case, then it might not have had the biasing influence over the deliberative mechanism that it did. Therefore, we would have to say that the recruiter is not responsible for his actions in this scenario.

Secondly, I consider the future possibilities for the job recruiter. Having done the implicit attitudes test, the recruiter is now aware of the existence of his implicit attitude as well as aware of its influence on his action. With this awareness the recruiter is now able to implement ecological control over his environment to shape his implicit attitudes and make them more reasons-responsive. Because of this, and because of the way this account of responsibility is structured, there are significant benefits to the way we can regard the recruiter's future behaviour. To demonstrate, I will take the original scenario but make three minor twists.

In the first scenario, having taken the implicit attitude test, the job recruiter is aware of his implicit attitude and its potential to influence his actions. To that end, he has implemented ecological control over his environment and has successfully managed to develop the reasons-responsiveness of his implicit attitudes. His attitudes now function in a very reasons-responsive way with his other evidence, and he consequently makes the informed choice, selecting the female candidate. As he has acted on a reasons-responsive mechanism, the job recruiter is morally responsible for this selection.

In the second scenario, having taken the implicit attitude test, the job recruiter is aware of his implicit attitude and its potential to influence his actions. Despite this he has decided not to carry out ecological control in an attempt to manipulate the reasons-responsiveness of his attitudes. When presented with the same scenario again, the job recruiter picks the male candidate; however, because there has been no change to his implicit attitude, he is still operating from a deliberative mechanism that is not completely reasons-responsive. Normally this would mean that the agent is not responsible for the action in question. However, because the agent is aware of his attitude and its potential influence, and has failed to implement ecological control, I argue that he is in fact morally responsible for his action (in virtue of his inaction).

In the third scenario, the job recruiter has again implemented ecological control over his environment, and has put in a lot of effort to trying to shape his attitudes such that he is not affected by them in his job. Nonetheless, the recruiter is faced with the same scenario and, despite his best efforts, he again selects the male candidate. In this case, because the agent has attempted to manipulate his environment, but has clearly not done so effectively enough, I argue that we still ought to hold him morally responsible for his actions, however I digress and assert that he may not necessarily be *blameworthy* (provided he has actually made a conceited effort to manipulate his attitudes).[99]

What I hope to have shown with this is how exactly we can apply this project in a practical setting. Not only have I demonstrated how we can evaluate the extent to which a person's deliberative mechanism is reasons-responsive, but also I believe I have shown what implications these conclusions have for instances in the real world. The account of moral responsibility I develop here, based on the Fischer and Ravizza (1998) account is evidently plausible and practical.

[99] I do acknowledge that an alternative perspective here would be that because the implicit attitude was so resistant to manipulation through ecological control, despite the agent's best efforts, that this is sufficient evidence that the attitude in question is not reasons-responsive. Therefore, as the agent has done what he could to pursue the most morally desirable outcome, and yet the attitude remained unchanged, that he is operating on a well and truly non-reasons-responsive mechanism, and therefore not morally responsible.

Having addressed the reasons-responsiveness criteria, I wish to briefly address another point; the criterion of guidance control. In chapter one I noted that an agent is not morally responsible for his actions if his actions issue from a *genuinely irresistible psychological impulse*[100] because it undermines guidance control. These irresistible impulses are described as being 'psychologically compelling' (Fischer & Ravizza, 1998). Given the automaticity of implicit attitudes, and how agents can sometimes lack any direct and immediate control to inhibit implicit attitudes, one might mount the argument that implicit attitudes are a kind of irresistible psychological impulse. In virtue of this, Fischer and Ravizza (1998) argue that the agent is not morally responsible.

I argue this would be incorrect. While implicit attitudes are indeed automatic and we lack direct and immediate control over them, we have demonstrated that they are far from irresistible. Given ecological control and the right motivation agent can inhibit or resist the influence and expression of implicit attitudes. If implicit attitudes functioned more like associations than patchy endorsements then the appeal to irresistible psychological impulses might carry more weight.[101] However as implicit attitudes are more reasons-responsive than this, I argue that they should not be regarded as irresistible psychological impulses, and so do not undermine the guidance control condition, and so do not absolve the agent of moral responsibility.

Finally, I wish to return to briefly discuss the ownership condition that I raised in chapter one, and discuss whether, on grounds of a lack of ownership, we can claim an agent affected by implicit attitudes is not responsible for their actions. There are a few points I want to make here. Firstly, as I have argued earlier in this project, implicit attitudes themselves are not a deliberative mechanism, rather they play a role *in* the deliberative mechanism of the agent. As such, I do not

[100] I also listed hypnosis, direct brain manipulation, brainwashing, and indoctrination as other examples which undermine guidance control. Again, this is not an all-inclusive list.

[101] After all, if implicit attitudes were associations then they would truly be automatic, unresponsive, and irresistible

believe it is the case that Fischer and Ravizza (1998) would require that the agent take ownership of their implicit attitudes *in the same way* as they are required to take ownership of their deliberative mechanism. However, this does not mean that we cannot have ownership over our implicit attitudes, and given that they can feature so significantly in the deliberative mechanism, I believe it is important for the success of this project that we do take ownership for these attitudes. But can we?

Fischer and Ravizza (1998) argue that when an agent takes responsibility for their deliberative mechanism (in an effort to meet the ownership condition), that agent takes responsibility for it in its full reality. In other words, we take responsible for the mechanism and all its details, including those which we are not aware of.[102] Therefore, in regards to unconscious mental states such as our implicit attitudes, Fischer and Ravizza argue that, even if we are not aware of them, we take responsibility for these states in virtue of taking responsibility for the deliberative mechanism. Therefore, I conclude that it is the case that so long as the agent is taking responsibility for the deliberative mechanism in question[103] then the agent also necessarily takes ownership of their implicit attitudes.[104] Thus, the agent satisfies the ownership conditions for

[102] To use Fischer and Ravizza's words, "we take responsibility for the whole iceberg in virtue of seeing its tip" (Fischer & Ravizza, 1998, p. 234

[103] Fischer and Ravizza mention in their 1998 text that there is virtually no incentive for the agent to not take responsibility for their deliberative mechanism (under normal circumstances), and that by failing to take responsibility for their deliberative mechanism, that the agent may suffer consequences much the same as if they had done so (Fischer & Ravizza, 1998, p. 239, footnote 32).

[104] One might question whether, given that our implicit attitudes can sometimes conflict with our explicit attitudes, this means that the implicit attitudes are not our own. Firstly, this claim seems to assume that our explicit attitudes are in fact our 'real' attitude (of which there is conflicting evidence). Secondly, assuming a dual construct model, even if our implicit attitudes are not our real attitude, they are nonetheless attitudes which we hold. In this case, we may not necessarily endorse these implicit attitudes, but they are nonetheless attitudes which are present in our deliberative mechanism in some way, even though they conflict with our explicit attitudes (this I believe is even more so the case on a single construct model). On the basis that we take responsibility (and ownership) for the components in virtue of taking responsibility (and ownership) of the mechanism, whether we endorse these attitudes or not does not matter; the fact remains that they are part of the mechanism and thus we have taken responsibility for them.

both the deliberative mechanism as well as the implicit attitudes, and so this cannot be grounds upon which to claim a lack of moral responsibility.

What I hope to have shown here is that actions which manifest our implicit attitudes can be eligible for moral responsibility. While determining the extent to which an agent is morally responsible for such behaviour requires examining the extent to which their implicit attitude is reasons-responsive, it remains possible for such behaviour to satisfy Fischer and Ravizza's (1998) criteria for moral responsibility. By regarding implicit attitudes as patchy endorsements, and by adopting ecological control, Fischer and Ravizza provide a way to hold agents morally responsible for their behaviours which manifest their implicit attitudes.

Conclusion

At the outset of this project I established three primary goals. Firstly, I wanted to provide an examination of the Fischer and Ravizza (1998) account of moral responsibility and examine its validity and applicability in a practical setting. I remarked that one of the most endearing features of the Fischer and Ravizza account is its realistic requirements and practical application. This theory sets reasonable standards for responsibility, and it provides an account of moral responsibility which can be adopted by determinists. Thus my first goal was to examine Fischer and Ravizza's theory.

I believe I have managed to provide such an examination. In chapter one I set out the challenge to moral responsibility if we define it by an agent having the ability to act otherwise than they do (thereby enacting the Principle of Alternate Possibilities). I then outlined and explained Fischer and Ravizza's (1998) distinction between guidance and regulative control, demonstrating that being able to 'do otherwise' is not necessary for moral responsibility. I extended upon this by providing an inquiry in to Fischer and Ravizza's notion of a moderately reasons-responsive deliberative mechanism. Having done so, and having established through this the proper requirements for moral responsibility, I addressed a few criticisms to the theory.

The second goal of this project was to perform an inquiry into the nature of implicit attitudes and their testing procedures. Despite considerable research gains into implicit attitudes in recent years, these psychological constructs are still far from being clearly understood. Therefore, throughout chapters two and three I endeavoured to provide a survey of the literature surrounding implicit attitudes and highlight what I believed to be key characteristics, as well as key misconceptions and challenges regarding their nature and operation. I argued that implicit attitudes are a psychological construct which display traits of automaticity, are beyond the agent's ability to directly and immediately control, and are known to be able to affect behaviour;

sometimes through interaction with explicit attitudes. I also examined popular testing procedures such as the IAT and priming procedures, and concluded that results of these methodologies provide ambiguous evidence, which may be interpreted as dual-process or single construct models of implicit attitudes.

It was at this point which I established my third goal, and that was to apply the Fischer and Ravizza (1998) account of moral responsibility to implicit attitudes. My goal was to see how Fischer and Ravizza's moral theory regarded behaviours which manifest implicit attitudes, and I pursued these goals over chapters three and four. I established the degree to which implicit attitudes are reasons responsive (arguing that implicit attitudes are more than mere associations but are not entirely bona fide beliefs), and chose instead to endorse a middle-ground alternative and regard them as patchy endorsements.

I carried this through to chapter four, in which I directly applied the Fischer and Ravizza (1998) theory to the understanding of implicit attitudes I had developed throughout chapters two and three. I determined that, if we were to adopt the Fischer and Ravizza account of moral responsibility, determining the extent to which an agent is morally responsible for behaviour manifesting implicit attitudes requires examining the extent to which their implicit attitudes are reasons-responsive. I concluded by noting that, if we consider implicit attitudes to be patchy endorsements, and we apply the notion of ecological control, then the Fischer and Ravizza account provides a way to hold agents morally responsible for their behaviours which manifest their implicit attitudes.

This project established three goals at the outset and I believe that I have managed to address each one in sufficient detail. However, while I have addressed a number of questions surrounding Fischer and Ravizza's (1998) moral theory, and implicit attitudes, my research has also brought to light many more questions and concerns which I believe need to be considered by future research, particularly regarding implicit attitudes.

First of all, are implicit and explicit attitudes single or dual constructs? In chapter two I briefly addressed this question by providing an overview of the prevailing theories, but assumed for the sake of this project, that implicit and explicit attitudes are dual-processes. However, due to the ambiguous nature of the evidence produced by implicit testing procedures, a firm and correct answer is still unclear. Thus future research in to implicit attitudes needs to provide a deep analysis in to both independent and dual construct theories in an effort to determine a solution. Doing so will have implications for implicit and explicit testing procedures; are we testing different constructs or just the one construct two different ways? There will also be implications for moral philosophy because if implicit and explicit attitudes are found to be a single construct (but still with some observed divergence) then the existence of elements which we may not previously have been aware of may make ascribing responsibility difficult.

Other directions to be taken by future research include determining to what extent we are morally responsible for *having* implicit attitudes. This project assumed that agents already have implicit attitudes and asked the degree to which we are responsible for the actions which issue from, or manifest, these attitudes. However, there is still the question regarding to what extent we are morally responsible for having these attitudes in the first place. I believe responses to this question would have to be formulated carefully, as it may have implications for questions such as ‘to what degree are we responsible for our environment’ or ‘to what degree are we responsible for our character’.

As a result of the size of this project I have been unable to address every possible contention or challenge that might have been raised by my research. For instance, in chapter four I claim that we can hold an agent morally responsible for failing to take ecological control if he is aware of implicit attitudes and their potential influence. Of course this raises a few challenges in itself, namely, does this mean we can only hold agents morally responsible as long as they are aware of their implicit attitudes? This seems like a very steep requirement to place on agents who may not know they have implicit attitudes, and especially for those agents who do not know what implicit attitudes are. How can we expect an agent to seek out an understanding of something which she does not know exists? At the same time, failing to do everything we can to understand our implicit attitudes may affect the degree to which we are morally responsible for their influence. Therefore, another question to be pursued by future research ought to be to what extent are we morally obligated to discover our implicit attitudes?

Even if the answer to this question is ‘yes we are’, responding in this way results in a run on effect; if we are morally obligated to learn about our potentially biasing implicit attitudes, are we morally obligated to learn about all potentially biasing cognitions we may be subject to? The answers one might respond with to these epistemic challenges may place limitations on the extent to which we can both ascribe moral responsibility on this account, as well as affect the way we talk about moral responsibility more generally and its relation to implicit attitudes. Again, due to the limited space in an already large project, I have not been able to address all of these concerns here. However, these are prevalent and important questions and I believe it is of the utmost importance that such challenges be taken up by future research projects.

The account of moral responsibility, and implicit attitudes, that I have constructed here provides epistemic and theoretical benefits, but also social and real world benefits. We live in a world at the moment where sexism, racism, religious prejudice, and sexuality discrimination are all very prevalent issues. I believe it safe to assume that, due to a variety of factors, almost every single person will possess some form of implicit attitude in regards to these topics, and oftentimes (but not always) these attitudes will have a negative valence. Instances of discrimination and prejudice can be found in the social sector, the business sector, the education sector, the medical sector, private life, and even in political agendas. Societies today are now starting to make a big push towards equality, by both trying to educate and change the way people implicitly respond to these issues. For the sake of equality, it is important that these implicit biases, discriminations, and misconceptions that we harbor be structures which can be adjusted in the face of reason and rationalization.

I believe the account of implicit attitudes that I have constructed here makes such goals a possibility. On my account, the manifestation and influence of implicit attitudes can be manipulated, changed, and inhibited. I have argued for a model of implicit attitudes which make them susceptible to rationalization and behaviour modification. Consequently, the account I have presented suggests that convincing, and logical (albeit extensive) arguments in favour of more egalitarian behaviours and values are ways in which we might be able to minimize the influence of implicit attitudes. If implicit attitudes were mere associations, then there would be little to no point in discussing gender equality or racial equality because associations simply do not respond to rationalization. By taking implicit attitudes to be patchy endorsements, and by adopting ecological control, we can actually justify opening dialogues about these issues and expect to observe changes in people’s behaviour.

While it is still the case that changing these behaviours would be a long and challenging process, the fact that it is a difficult process does not mean we are not morally bound to carry it out. And while it is the case that some of these changes to behaviour may not always be long-lasting, the fact that it is possible to affect these behaviours at all is enough to encourage us to try. It serves to motivate us to act in such ways as to promote equality, and to hold those people who behave in prejudicial and discriminatory ways due to implicit attitudes morally responsible for their actions. If we want to be able to hold these people responsible for their discriminatory behaviour (and I believe we should be able to), then I believe Fischer and Ravizza's (1998) account of moral responsibility provides us this ability.

References

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 3*, 242-261.

Clark, A. (2007). Soft selves and ecological control. In Spurrett, D., Ross, D., Kincaid, R. H., & Stephens, L. (Eds.), *Distributed cognition and the will*. Cambridge, MA: The MIT Press.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12,* 163-170.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Science, 17,* 363-366.

Dasgupta, N., & Greenwald, A. (2001). On the malleability of automatic attitudes: Combining automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800-814.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8/7,* 342-353.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *21*, 1139-1150.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33,* 510-540.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54,* 297-327.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*, 829-839.

Frankish, K. (2015). Playing double: Implicit bias, dual levels, and self-control. In Brownstein, M., & Saul, J. (Eds.), *Implicit bias and philosophy volume 1: Metaphysics and epistemology*. Oxford University Press.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology, 108*, 553-561.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellot, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*, 3-25.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit attitude test. *Journal of Personality and Social Psychology, 6*, 1464-1480.

Hoefer, C. (2010). Causal Determinism. *Stanford Encyclopaedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/determinism-causal/

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations, 11*, 69-87.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*, 274-306.

Holroyd, J., & Kelly, D. (forthcoming). Implicit bias, character, and control. In Webber, W., & Kelly, D. (Eds.), *From personality to virtue*.

Hughes, S., Barnes-Holmes, D., & Holroyd, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioural alternatives. *The Psychological Record, 61*, 465-496.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit associations test. *Journal of Personality and Social Psychology, 81*, 774-788.

Levy, N. (2014). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Nous*, 1-24.

Madva, A. (2012). *The hidden mechanisms of prejudice: Implicit bias & interpersonal fluency* (Doctoral dissertation). Columbia University.

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, 1-36.

McKenna, M. & Coates, D. J. (2015). Compatibilism. *Stanford Encyclopaedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/compatibilism/

Mele, A. R. (2000). Reactive attitudes, reactivity, and omissions. *Philosophy and Phenomenological Research, 61*, 447-452.

Payne, B. K., Cheng, C. M., Govorun, C., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277-293.

Pereboom, D. (2006). Reasons-responsiveness, alternative possibilities, and manipulation arguments against compatibilism: Reflections on John Martin Fischer's *My Way*. *Philosophical Books, 47*, 198-212.

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology*, *44*, 29-45.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17,* 954-958.

Rydell, R., & McConnell, A. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*, 995-1008.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101-126.